Universidade de Aveiro Departamento de
2015 Eletrónica, Telecomunicações e Informática

Renato Pinho Rodrigues

Métodos baseados em grafos para desambiguação de conceitos biomédicos

Graph-based methods for biomedical concept disambiguation


Universidade de Aveiro
2015
Departamento de
Eletrónica, Telecomunicações e Informática

Renato
Pinho Rodrigues

Métodos baseados em grafos para desambiguação de conceitos biomédicos

Graph-based methods for biomedical concept disambiguation

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Doutor Sérgio Aleixo Matos, Investigador Auxiliar do Instituto de Engenharia Eletrónica e Telemática de Aveiro

o júri / the jury

Presidente / president	Prof. Doutor Joaquim Manuel Henriques de Sousa Pinto Professor auxiliar da Universidade de Aveiro
Vogais / examiners committee	Prof. Doutor António Manuel de Jesus Pereira Professor coordenador do Instituto Politécnico de Leiria
	Doutor Sérgio Guilherme Aleixo de Matos Investigador auxiliar da Universidade de Aveiro

agradecimentos / acknowledgements

Gostava de agradecer em primeiro lugar ao meu orientador Sérgio Matos pela orientação, apoio e paciência infinita durante a realização do mestrado. Todas as experiências partilhadas, ensinamentos transmitidos e discussões foram fundamentais para a realização deste trabalho. Gostava também de agradecer ao Professor José Luís Oliveira pela oportunidade de integrar o Grupo de Bioinformática durante o mestrado, bem como por todas as conversas e ensinamentos, e ainda agradecer a todos os membros do Grupo por toda a disponibilidade e conselhos que tanto contribuíram para a minha formação académica e pessoal.

Também quero agradecer aos meus amigos Anton Zverev, Jóni Lourenço e Michael Domingues pelos anos de aventuras, crescimento e aprendizagem ao longo do curso, bem como ao Luís Ribeiro, David Campos, Luís Lemos e Luís Fernandes pelas excelentes conversas, enorme paciência, apoio, e momentos de diversão. Aproveito ainda para agradecer também a todos aqueles que ao longo destes anos contribuíram direta ou indiretamente para a minha formação académica e pessoal, nomeadamente aos meus colegas do curso Mestrado Integrado em Engenharia de Computadores e Telemática e aos colegas com quem trabalhei na Associação Académica da Universidade de Aveiro.

Um agradecimento especial à minha família, aos meus avós António e Judite, à minha mãe Ana e ao Fernando, aos meus tios António José, Graciete, Rui e Elsa, pelo apoio ilimitado durante toda a minha formação, bem como pelo amor, amizade e educação que sempre me deram. Ainda um agradecimento ao meu pai Jorge e à Carla, também pelo apoio e amizade ao longo dos anos.

Por último mas não menos importante, quero agradecer à minha namorada Bárbara Pedro pela amizade, paciência e pelo apoio extraordinário que tantas vezes tornou mais fácil superar grandes desafios.

palavras-chave

Bioinformática, mineração de texto, extração de informação, desambiguação de conceitos, reconhecimento de conceitos.

resumo

Desambiguação do sentido das palavras é a tarefa de atribuir um significado inequívoco a uma palavra ou termo ambíguo, tendo em conta o contexto em que este está inserido. O domínio da biomedicina contem um grande número de termos ambíguos, não identificar corretamente o sentido associado a cada termo tem um impacto negativo na performance de aplicações biomédicas tais como as de anotação automática e indexação, as quais são cada vez mais de extrema importância no contexto biomédico e clinico, dado o rápido crescimento da informação digital disponível para os investigadores.

Este tese foca-se na desambiguação de termos biomédicos e apresenta uma solução que atribui identificadores únicos a palavras ambíguas baseando-se, para isso, no Unified Medical Language System (UMLS). O método proposto é uma aproximação baseada em fontes de conhecimento a qual não necessita de dados de treino, sendo assim uma solução generalizada que pode ser amplamente aplicada para resolver ambiguidades no domínio biomédico. Este método baseia-se em grafos obtidos a partir do UMLS, tendo em consideração os conceitos presentes no contexto da palavra ambígua, e utiliza um algoritmo de PageRank para atribuir pontuações aos grafos. Adicionalmente foi desenvolvido e disponibilizado um web-service para uma fácil integração em aplicações de terceiros, com o objetivo de munir essas aplicações com um módulo fácil de usar e com grande potencial.

O sistema foi testado e avaliado utilizando uma coleção de testes de desambiguação de conceitos, desenvolvido pelo U.S. National Library of Medicine, especificamente o *MSH WSD Test Collection*, um conjunto de dados que contém mais de 37 mil ocorrências de 203 termos ambíguos.

Os melhores resultados obtidos pelo sistema proposto alcançaram uma precisão de 63.3% no *subset* do *MSH WSD Test Collection*.

keywords

Bioinformatics, text mining, information extraction, word sense disambiguation, knowledge bases, concept recognition.

abstract

Word Sense Disambiguation (WSD) is the task of assigning a unique meaning to an ambiguous word or term, given the specific context it is inserted in. The biomedical field contains a large number of ambiguous terms, and not being able to correctly identify the correct sense associated to a term has a negative impact on the accuracy of biomedical applications such as automatic annotation and indexing, which are becoming of utmost importance in the biomedical and clinical world given the fast growing amount of digital information available to researchers.

This thesis focuses on disambiguation of biomedical terms and presents a solution that can assign unique identifiers to target words based on Unified Medical Language System (UMLS). The method proposed is a knowledge-based approach where no training data is required, thus being a more general solution that can be widely applied to solve ambiguities in the biomedical domain. This method relies on graphs obtained from the UMLS, taking into consideration the concepts from the context of the ambiguous word, and uses a PageRank algorithm to score such graphs. Furthermore a web-service was developed and made available for an easy integration in third-party applications, in order to provide such applications with a powerful and easy to use module.

The system was tested and evaluated using a WSD test collection provided by the U.S. National Library of Medicine, specifically the MSH WSD Test Collection, a dataset containing over 37 thousand occurrences of 203 ambiguous terms.

The best performing results of the proposed system achieve an accuracy of 63.3% for a subset of the MSH WSD Test Collection.

Contents

List of Figures

List of Tables

Acronyms

AEC Automatic Extracted Corpus

API Application Programmatic Interface

ARFF Attribute-Relation File Format

AUI Atom Unique Identifier

BFS Breadth First Search

CUI Concept Unique Identifier

HITS Hypertext Induced Topic Selection

JDI Journal Descriptor Indexing

MeSH Medical Subject Headings

ML Machine Learning

MRD Machine Readable Dictionary

PPR Personalized Page Rank

NLM National Library of Medicine

UMLS Unified Medical Language System

WSD Word Sense Disambiguation

Chapter 1

Introduction

The constant growth of available data has originated a lot of interest in automated tools that can obtain and assess this information effectively [1], either for research or industrial purposes. Across all fields data is being collected and stored rapidly. However, a large portion of this information is available in the form of free text creating several challenges regarding its direct use in computerized solutions. Text mining is a computational area that addresses these challenges through autonomous techniques, and that is being explored by big IT companies, such as Google, Amazon, IBM or Microsoft and also by its potential clients, from the financial area to the pharmaceutical industry.

With this constant increase of information, recorded in form of free text, a lot of effort is put into developing efficient techniques that can identify, extract, manage, integrate, and exploit it. Text Mining is the field that deals with those requirements, by deriving high-quality information from text. The main goal of this field is to retrieve information and represent it in structured form, thus enabling its use to induce knowledge through combination of several sources [2]. In order to do this, it is necessary to link data to a specific domain and field which, in some cases, might not be an easy task.

The biomedical domain is divided in many fields, existing various relation type between concepts from different fields (for example, diseases are often related to genes). Furthermore, this domain is growing and evolving constantly as new concepts and knowledge is developed almost every day. In addition, the domain specific and non-standard terminology results in high levels of ambiguity as the same terms are constantly used with different meanings and regarding different fields of the biomedical domain.

Therefore, the development of text mining solutions in this domain is a major challenge due to its range and complexity.

Several information extraction methods and systems have already been developed taking into consideration the requirements, evaluation strategies and tasks that need to be performed to accomplish the information extraction goals. These tasks were introduced by the Message Understanding Conferences [3] and include:

- Named Entity Recognition: identify specific entity names, such as people and organization, in text;
- Normalization and disambiguation: associate an unique meaning to a concept (e.g. “dare” could refer to the English verb, a protein or an organism);
- Coreference: identify occurrences of two different expressions that refer to the same concept;
- Relation Mining: extract relations between concepts;
- Summarization: extract and compile main ideas of a text;
- Classification: identify prime themes of a specific text.

To efficiently complete the information extraction pipeline is a complex process as some of its tasks may be challenging. In this thesis, the disambiguation task will be focused and further explored.

Ambiguity, i.e., words with multiple meaning or senses, is very common in natural languages. For example, the word “watch” in the sentence “Tom bought a new watch” refers to the small timepiece worn on one’s wrist, whereas in the sentence “Lucy kept him under watch” refers to the act of observing someone over a period of time. The correct sense of the word “watch” can be obtained from its context. The context in which ambiguous terms are inserted helps in disambiguating and understanding the correct meaning of the word. Nevertheless, while this process may appear to be trivial for a human, for computers that is not the case. In fact, even humans sometimes disagree on the interpretation of the senses of terms. Furthermore, terms may have more than two meanings, which increases the difficulty of the disambiguation task. For example, the word “take” is listed in the Oxford Dictionary with nine different senses as a verb and another two as a noun.

In the biomedical domain, ambiguity may be more common than in general language. For example, it is frequent that a gene, a protein encoded by the gene and a disease associated with the protein have the same name. In such scenario, it is important that the correct meaning of the term is identified in order to associate a correct identifier (either gene, protein or disease) to the term. However, the different meanings of the term depend on its context and cannot be induced looking at the word itself. A correct analysis of the context is required in order to disambiguate the concept.

In Word Sense Disambiguation (WSD), meanings are known as senses, which are obtained from dictionaries or other lexical resources. Ambiguous words are referred to as target words and their context is known as an instance. The instance can be a phrase, sentence or a paragraph. The instance can in some cases, if needed, be expanded to the whole document.

The main goal of this thesis is to propose a solution for word sense disambiguation problem, using the Unified Medical Language System (UMLS) as a knowledge source and implementing unsupervised algorithms i.e., algorithms that do not require any training data or human interaction.

The contributions of this thesis are:

- The research of biomedical word sense disambiguation techniques that do not rely on manually annotated data;
- The development of an extensible, scalable and easy to integrate tool for biomedical word sense disambiguation.

The remainder of this thesis is organized as follows:

- Chapter 2 presents an analysis of the current work in biomedical word sense disambiguation problem, introducing some key concepts needed to understand WSD and giving an overview of the existing solutions and overall achieved results.
- Chapter 3 discusses the proposed system. Firstly the motivation behind the selected approach is explained. Secondly the overall architecture and implementation details are explored. Lastly, the scoring technique used in the system is discussed.

- Chapter 4 contains the overview and explanation of the results obtained from the experiments conducted using a manually annotated dataset as the validation of the system.
- Chapter 5 discusses some potential future work in biomedical WSD, specifically some proposals to improve this system's performance.
- Chapter 6 presents the contributions and overall conclusions of this thesis.

Chapter 2

Word Sense Disambiguation

Word Sense Disambiguation is the task of associating a unique meaning to an ambiguous term, given a specific context. This task is usually simple for humans, however for machines it is a quite challenging task due to the necessity to understand the context in which the word is inserted. This process usually not only involves analyzing the preceding and following words of the term to disambiguate, but it is also necessary to have background knowledge on the different senses of the word.

Term ambiguity is very common in the biomedical field because of the complexity of the domain. Jimeno-Yepes and Aronson [4] conducted an analysis of MEDLINE and concluded that the term *study* is mapped to six different concepts more than three million times, being the most ambiguous term. They also concluded that the majority of ambiguous concepts belong to “Gene or Genome” and “Amino Acid, Peptide, or Protein” semantic types from UMLS. On a different analysis, Weeber et al. [5] concluded that 11.7% of sentences present in MEDLINE were ambiguous relative to the UMLS Metathesaurus.

When WSD is successful, it increases the number of biomedical terms normalized correctly, i.e. associated to the correct ontology or database concept, thus helping improve application that are widely used by researchers, such as automatic annotation and indexing, information extraction and knowledge discovery. For instance, Aronson [6] states that MetaMap, a program that maps biomedical text to UMLS concepts initially developed to improve retrieval of bibliographic material such as MEDLINE citations, would be improved if a WSD component were present. This component was later integrated in MetaMap, as presented by Aronson et al. in 2010 [7]. WSD solutions require the

application of advanced disambiguation techniques, which are not trivial and require a large amount of curated knowledge.

Given the above-mentioned facts it is easy to understand that correctly assigning senses to terms is important. However, in order to do so, it is necessary to train a machine to acquire the correct knowledge to perform such distinction and make the machine aware of the multiple senses and meanings of terms. Regarding the first part, i.e., training the machine to distinguish senses of a term, it is an accepted notion that WSD is an AI problem as difficult as any other [8]. Considering the difficulty of the task and its importance much work has been done in this area. Some of the most popular algorithms are further explored in this chapter (see Section 2.1, 2.2, 2.3).

In respect to making the machine aware of multiple senses of terms, the concept of knowledge base appears. The most popular knowledge base for WSD is WordNet[9], a large lexical database of English that stores nouns, verbs, adjectives and adverbs in synonym sets (synsets), each defining a concept. In addition, WordNet also stores semantic relations between synsets, establishing networks of related words.

In the biomedical domain, the most used knowledge base is the UMLS, specifically the UMLS Metathesaurus and UMLS Semantic Network. The Metathesaurus contains millions of biomedical and health related concepts, as well as their synonyms and relations, while the Semantic Network provides a categorization of such concepts and a set of semantic relations between semantic types. The Semantic Network contains 133 semantic types and 54 semantic relationships, where the major semantic types include “organism”, “anatomical structure” among others, and “part_of”, “treats” and “interacts_with” are some examples of relations. Figure 1 illustrates a small subset of the semantic network semantic types hierarchy.

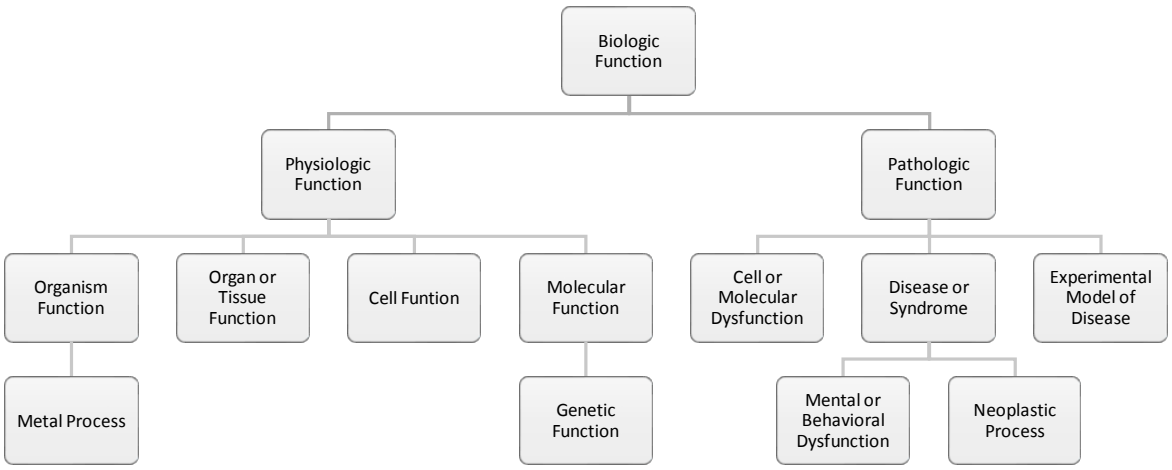

Figure 1: Example of semantic types in UMLS semantic network

In general terms, current solutions for WSD can be categorized as being machine learning based or knowledge-based. ML-based solutions apply techniques to automatically learn which concept is associated with a specific term. Such solutions can be further divided as supervised, semi-supervised and unsupervised based on their learning method. Supervised and semi-supervised approaches train a statistical model to classify target terms and assign a unique concept based on its context. Due to the nature of those two methods, they will be addressed as classification-based ahead in this thesis. On the other hand, unsupervised algorithms apply clustering techniques to build document clusters related to the terms to disambiguate. For this reason, unsupervised approaches are categorized as clustering-based in this document.

2.1. Resources for WSD

2.1.1. Knowledge Sources

Several institutions created standards for concept names definition in order to provide centralized resources and motivate their association with patients' health records

and research laboratory resources. Nevertheless, despite the success of the standardization processes, several biological concepts still lack standards for careful names definition. Moreover, all variant names of a specific concept are not contained in a single resource. Therefore it is important to combine the different existing knowledge bases in order to collect as much information as possible regarding a specific concept type.

Table 2.1 presents a list of publicly available biomedical databases and ontologies which may contain relevant data for concept recognition and disambiguation.

Table 2.1: Biomedical databases and ontologies

	Name	Concept(s)
Databases	Entrez Gene	• Gene
	HUGO Gene Nomenclature Committee (HGNC)	• Gene
	Uniprot	• Protein
	Protein Data Bank (PDB)	• Protein
	Expert Protein Analysis System (ExPASy)	• Enzyme
	ChemIDPlus	• Chemical
	Humam Metalobome Database (HMDB)	• Small molecules
	DrugBank	• Drug
	Pharmacogenomics Knowledge Base (PharmKGB)	• Gene • Drug • Disease
	RxNorm	• Drug
	Kyoto Encyclopedia of Genes and Genomes (KEGG)	• Pathway
	BioSystems	• Pathway
	Online Mendelian Inheritance in Man (OMIM)	• Disease • Variation • Gene
	Systematized Nomenclature of Medicine	• Anatomy

	(SNOMED)	• Morphology • Species • Chemical • Drug • Disease • Diagnosis • Procedure • Physical agents, forces, activities • Social context
	Medical Subject Headings (MeSH)	• Protein • Chemical • Disease
	Comparative Taxogenomics Database (CTD)	• Gene • Chemical • Disease • Pathway
	Medical Dictionary for Regulatory Activities	• Disease
Ontologies	Chemical Entities of Biological Interest (ChEBI)	• Chemical
	Cell Ontology (CL)	• Cell
	Gene Ontology (GO)	• Gene
	Protein Ontology (PRO)	• Protein
	Sequence Ontology (SO)	• Sequence
	Disease Ontology (DO)	• Disease
	National Center for Biotechnology Information (NCBI) taxonomy	• Species
	Common Anatomy Reference Ontology (CARO)	• Anatomy
	Unified Medical Language System (UMLS) semantic-network	• Species • Anatomy • Chemical

		• Biological function • Physical object • Idea or concept

The Unified Medical Language System (UMLS)[10] is one of the most used knowledge bases in WSD applications, because of its large spectrum of concepts and because it also provides semantic relationships. Having the ability to have several concept types agglomerated in a single knowledge base brings advantages for WSD solutions, thus some researchers put their efforts into developing aggregators that afterwards can be used as knowledge sources (e.g. Linked Life Data).

UMLS is a repository developed by the US National Library of Medicine to support biomedical research. It currently aggregates over 100 source vocabularies[1] with different semantic and syntactic structures. The UMLS consists of three major components: the Metathesaurus, Semantic Network and SPECIALIST Lexicon.

The major component is the Metathesaurus, a multi-lingual repository of interrelated biomedical and health concepts from various sources. Some of which are semi-automatically integrated, such as Medical Subject Headings (MeSH) and SNOMED Clinical Terms (SNOMED-CT).

SNOMED-CT is one of the biggest sources in the UMLS Metathesaurus. This source was originally created by the College of American Pathologists with the purpose of improving patient care by developing systems to record health care encounters accurately. It includes more than 300,000 unique concepts with 903,000 links or semantic relationships. MeSH was developed and is maintained by the U.S. National Library of Medicine (NLM). It is a vocabulary used for indexing, cataloging and searching biomedical and health-related information. MeSH has more than 25,000 concepts in the Metathesaurus, arranged in a hierarchical structure.

The Metathesaurus organizes knowledge based on Concept Unique Identifiers (CUI), containing over 3 million CUIs and more than 35 million relationships between them.

[1] Full list of vocabularies available at:
http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

Each CUI has a set of specific attributes, such as: preferred term, concept definition, associated terms, related concepts. The concepts can be related to each other and there are 12 different types of relations that can exist between concepts. Table 2.2 shows a full list of relationships and their description.

Table 2.2: List of UMLS relationships and meaning

PAR / CHD	Has parent / child relationship
RB / RN	Has broader / narrower relationship
AQ	Allowed qualifier
QB	Can be qualified by
RL	Concepts are similar or alike
RO	Has relationship other than synonymous, narrower or broader
RQ	Related and possibly synonymous
RU	Related, unspecified
SIB	Has sibling relationship
SY	Source asserted synonymy
XR	Not related
DEL	Deleted concept

The information about each CUI comes from one or more sources of information. A concept within a specific source is identified by an Atom Unique Identifier (AUI). All attributes of the AUIs are also associated with its corresponding CUI. Each CUI entry is also linked to the original source of information code, allowing the mapping to the original data. An example of the mappings between concepts (CUI), atoms (AUI) and original source unique identifier (Code) is shown in table 2.3.

Table 2.3: Example linkage between CUI, AUI and Code

Concept (CUI)	Atoms (AUI)	Code
C0007776 Cerebral Cortex (preferred) Structure of cerebral cortex	**A0036988** Cerebral Cortex	**D002540** (From MeSH)
	A10894985 Structure of cerebral cortex	**40146001** (From SNOMED-CT)

2.1.2. Corpora

Publicly available biomedical corpora for development and evaluation of WSD systems is very limited. Due to the complexity of the field, some authors have to create specific test data sets for their specific tasks. However, in the last years, with the growth in popularity of WSD systems, some general sets were built specifically for WSD. The following corpora are commonly used:

- NLM WSD test collection: uses the 1998 MEDLINE as baseline, containing 50 ambiguous words identified in 5 thousand MEDLINE citations;
- MSH WSD test collection: the latest version contains over 37 thousand MEDLINE abstracts annotated with 203 ambiguous terms with almost 38 thousand occurrences. Is one of the most popular data sets for WSD;
- Medstract: focused on acronym disambiguation, contains 173 acronym-meaning pairs;
- MuchMore: based on the Springer corpus of medical abstracts, contains both English and German versions of the same abstracts. All ambiguous words in the abstracts were annotated, however the inter-annotators agreement is considerably low, with 65% for German and 51% for English.

Nevertheless, other corpora developed for concept recognition and normalization may be used to evaluate WSD systems, since they provide unique concept identifiers for each named entity. The following table (Table 2.4) presents a list of some of the most relevant corpora for biomedical concept disambiguation considering the source of annotations and target concepts. Furthermore, only gold standard corpora providing unique identifiers from known knowledge bases are listed.

Table 2.4: List of relevant corpora

Corpus	Year	Concepts
BioCreative II GN	2008	• Gene and protein
BioCreative III GN	2011	• Gene and protein
OrganismTagger	2011	• Species
Linnaeus	2010	• Species
EBI Disease	2008	• Disorders
Arizona Disease	2009	• Disorders
CRAFT	2012	• Gene and protein • Species • Chemical • Cell • Biological processes • Molecular functions • Cellular components

2.2. Classification-based WSD

In these approaches the problem is addressed as a classification task: Classify an ambiguous term in a given context to one of its potential senses using any of the existing classification algorithms (decision trees, naïve Bayes or Support Vector Machines, for example). In general terms, classification-based WSD consists on collecting a set of documents that contain the ambiguous term such that the sense of the term is known for that specific context. From the set of documents, train a model for the given sense. Afterwards, given a new occurrence of the ambiguous term, the trained models for each sense are compared with the context and the sense whose model fits best is assigned. Such approaches take advantage of rich sets of features, such as:

- Linguistic: tokens, lemmas, POS, chunks and dependency parsing;
- Morphological: char n-grams and word shape;
- Local context: windows of features and/or conjunctions of tokens' features that co-occur frequently and that contain the ambiguous term;

- Distance and position: the position of the token to the ambiguous term, through its distance and orientation (left or right);
- Domain knowledge: recognize named entities, such as disorder, drugs and procedures;
- Document metadata: section heading and medical specialty.

As stated before, classification-based approaches consist of both supervised and semi-supervised learning techniques. Supervised techniques rely on a training data set to create a generalized model. Training data consists of inputs that are tagged with predefined class labels. The accuracy of such techniques depends on the quality and type of the data used to build the model. The training data must be representative of real-world conditions and be variable enough to ensure all conditions are captured in the generalized model.

Regarding semi-supervised approaches, the use of unlabeled data on biomedical WSD revealed a positive impact, applying both self-training and co-training [11]. These techniques use a smaller amount of training data as initial seed to train the first model for WSD. Afterwards, this model is applied to an untagged corpus and the terms classified with a high confidence rate are added to the initial training data. Then a new model is trained and the process is repeated until all terms are classified with high confidence.

Classification-based WSD usually achieves very good results and for that reason is the chosen method for some applications. However, both supervised and semi-supervised solutions are limited by the amount and quality of existing annotated data. Analyzing one of the largest and most used corpora, NLM WSD, it only provides 203 ambiguous terms, which is restrictive considering the complexity of the biomedical domain. Furthermore, “at least a few dozen” [12] labeled examples per ambiguous term are necessary to develop competitive classification-based applications. Therefore, a huge effort is required to generate such amount of data with acceptable quality. This limitation means that most supervised techniques can only be applied in a small scope, therefore limiting its usefulness in practice. For that reason, the interest for solutions that do not require training data has increased, with the application of clustering algorithms and knowledge based approaches.

Table 2.5 presents results obtained using supervised (Naïve Bays) and, semi-supervised (Automatic Extracted Corpus) techniques. In the first, words present in the citation where the ambiguous term appears are used as features in a supervised Naïve Bays algorithm. The results shown are based on 10-fold cross-validation. The latter method, Automatic Extracted Corpus (AEC), aims to reduce the problem of availability of manually annotated training data. For this system, training data used to train a machine learning algorithm is automatically generated using documents from MEDLINE, related to the possible senses of the ambiguous term.

Table 2.5: Classification-based algorithms performance

Method	NLM WSD Set	NLM MSH Set
Naïve Bays	0.8830	0.9386
Automatic Extracted Corpus	0.6836	0.8383

2.3. Clustering-based WSD

Clustering-based techniques (also known as unsupervised learning) is also a set of machine learning methodologies used to detect senses from examples rather than deciding the correct sense for a specific ambiguous term. Since these methods do not rely on preconceived class labels, no training data can be created to generate a model.

Given a set of examples of texts containing a particular term, the discovery technique attempts to find patterns between those and cluster them into groups of similar text. From these clusters, common labels can be selected. Any text clustering technique may be used and the distance measure between texts may be based on the cosine distance, the Jaccard coefficient or any other. As a result, it is expected to obtain as many clusters as senses for the given term, i.e., each cluster should contain only occurrences of the term with the same sense.

Some authors developed approaches which induce senses from untagged corpus. Those were known as unsupervised word sense induction [13], [14]. The main motivation for this approach comes from the observation that sometimes the definitions contained in lexical databases sometimes do not reflect the exact meaning of a term. Such solution

aims to overcome the limitations of lexical databases by extracting the different senses of a term from the corpus itself. Graph based algorithms which induce corpus senses by partitioning the co-occurrence graph of a term became popular under this approach. One drawback of these algorithms is the need for a large number of untagged instances for each term to induce relevant partitions in the co-occurrence graph.

The advantage of this method, when compared to classification-based methods, is that no training data is necessary and it is not necessary to predefine sets of senses for each term. Thus, the trade-off between clustering and classification approaches is the accuracy of the results versus the intensive human effort of developing training data. Moreover, clustering techniques are typically general and do not take advantage of domain knowledge.

Martín-Wanton et al. [15] present a method where concepts are represented as vectors built using the concepts from UMLS, containing the words of the concept definitions and it frequency. More examples of clustering-based methods are the Journal Descriptor Indexing (JDI) [16] and Machine Readable Dictionary (MRD) [17] approaches. JDI method aims to assign a concept to an ambiguous term by identifying its semantic type. Two types of vectors are created in this method. A vector for each semantic type of the possible concepts and a vector representing the term, containing the words that exist in the context, are compared and the concept whose semantic type vector is closest to the ambiguous term vector is assigned to the term. The MRD approach creates a vector for the ambiguous term using its context and a vector for each of its possible concepts. Afterwards, cosine similarity between concept vectors and the context vector is calculated, selecting the concept with highest cosine similarity. Table 2.6 presents the overall accuracy of these three methods, with the particularity that the JDI approach is not able to disambiguate terms whose possible meanings share the same semantic type, therefore excluding 44 terms from the data set.

Table 2.6: Clustering-based algorithms performance

Method	NLM MSH Set
Martín-Warton et. Al	0.7438
Machine Readable Dictionary	0.8070
Journal Descriptor Indexing	0.6551

2.4. Knowledge-based WSD

Knowledge-based approaches rely on the existence of knowledge sources, such as WordNet [18], Thesaurus[2], or the Unified Medical Language System (UMLS) [10] for the life sciences scope. Usually, such solutions do not use any information from corpora, therefore eliminating the need for high quality curated corpus, hence lowering the human effort drastically. For that reason, knowledge-based techniques have wider-applicability.

In the biomedical field, such approaches were firstly used to disambiguate gene and protein names, since a gene may have multiple species associated, and consequently as many concept unique identifiers. The idea behind such approach is to use external information to detect the correct unique identifier [19]. Taking in consideration that genes and proteins contain a large amount of related information on biomedical resources, such as diseases, functions, mutations and domains, for each identifier that is candidate for the ambiguous term, the method will find all information that is related with the gene or protein in the surrounding text, and the identifier with highest likelihood is selected.

On the other hand, for more specific and specialized solutions more complex approaches may be applied, taking advantage of machine-learning and/or fine-tuned filters. For instance, Liu, Johnson and Friedman [20] proposed a method using UMLS as the ontology and identifying UMLS concepts in abstracts. Afterwards they analyze the co-occurrence of these terms with the term to be disambiguated and build their word sense-tagged corpora automatically instead of manual annotation as most supervised techniques require.

Furthermore, some authors identify three types of knowledge-based approaches, considering graph-based approaches, approaches that use semantic similarity measures and overlap-based approaches. The later approaches, relies on the features describing the term to be disambiguated and the senses listed on the lexical database. Essentially, such approaches form bags-of-words containing the features of each listed sense (sense bags), capturing the behavior for each sense. Similarly, a bag-of-words is created with the features describing the term in the given context (context bag). Thereafter, the correct

[2] http://www.thesaurus.com/

sense of the term is identified by finding the maximum overlap between the sense bags and the context bag.

Regarding semantic similarity measures approaches, those have been developed to exploit and analyze the network of semantic connections between word senses. Such approaches were popular in the early 1990s with the appearance of WordNet. Some of the most popular were proposed by Agirre and Rigau [21], Leacock and Chodorow [22], Lin [23], and Resnik [24]. Basically these approaches aim to select the sense of a term which maximizes its semantic similarity with other words in the context. However, the proposed algorithms showed very low accuracy and failed to outperform the previous methods [8].

Finally, graph-based approaches are based on exploiting the graph structures to determine the appropriate sense for the given context. Firstly Mihalcea (2005) [25], then Agirre and Soroa (2009) [26] proposed the use of PageRank for finding the best combination of senses in a sense graph. The later authors, inspired their solution in the Google Page Rank algorithm, using it to encode word sense dependencies through random walks on graphs. In their approach, UMLS is represented as a graph, with concepts represented as vertices and relations between concepts as edges. For this algorithm, two sources of information are needed, a knowledge base and a dictionary to map words found in documents to their possible concepts in the knowledge base.

Recently, knowledge-based approaches started to be widely applied to perform biomedical disambiguation. In most cases, the knowledge resource used is UMLS Metathesaurus, since it provides a large coverage of biomedical domain knowledge and is constantly updated and maintained. Consequently, varied solutions have emerged in the last few years, using advanced scoring techniques to find the concept more related with the context of the ambiguous term.

Basic knowledge-based approaches are simple and easy to implement, as they rely on simple lookups on knowledge resources. Also, they do not depend on any type of corpus, either tagged or untagged, because no training process is needed. On the other hand, these solutions have poor accuracies. Jimeno-Yepes and Aronson [4] performed a study on knowledge-based approaches for biomedical WSD, comparing the achieved results in the NLM-WSD corpus. The authors showed that the Machine Readable Dictionary ap-

proach achieved an accuracy of 63.9%, outperforming the Personalized Page Rank (PPR) with 58.3% accuracy on the NLM WSD dataset. Nevertheless, the best results were obtained by combining three approaches – MRD, PPR and AEC – achieving a total accuracy of 76.3%.

These results are lower than results obtained by supervised classification approaches. However, the wider applicability of such approaches justifies their application. Existing tagged corpora in the biomedical domain cover a small number of terms and senses when compared to UMLS Metathesaurus. Since extending existing corpora is not feasible, the existence of techniques such as knowledge-based are of utmost importance.

The main downside of such approaches is that they require a complete dictionary. While it is possible to have such dictionaries for natural languages (e.g. the Oxford Dictionary), for Life Sciences specific terms this scenario is not as simple due to the large lexical and morphological variability of the terms. In such cases, resources such as UMLS are helpful as they also provide semantically close words for given senses.

McInnes et al. [27] present UMLS::SenseRelate, a WSD system that uses the degree of similarity between the possible senses of the ambiguous word and the terms present in its context. This method was evaluated using path-based [22], [28], [29], and information-content measures [24], [30], [31]. The first relies on hierarchical relations between terms, the latter enriches this information by quantifying the specificity of a concept in the hierarchy. Table 2.7 shows the best performing accuracy of this system using the path-based and the information-content measures.

Table 2.7: Knowledge-based algorithms performance

Method	NLM MSH Set
Path-based measure – Nguyen & Al-Mubaid [29]	0.72
Information-content measure – Lin [31]	0.74

Chapter 3

Requirements and Implementation

This chapter describes the proposed Word Sense Disambiguation method. An unsupervised, graph-based, technique was chosen to address the problem of WSD. This method does not require training data and merely depends of a knowledge source. UMLS was used as knowledge source for the purpose of this work, nevertheless any other database could be used in this method.

The following sections present the motivation of using a graph-based approach, the requirements of the system and its architecture, as well as the scoring algorithms used.

3.1. Motivation

Despite the fact that, generally, knowledge-based techniques have lower accuracy results than the remaining (see Chapter 2), it was the chosen approach to this system. The motivation behind this decision is the assumption that such techniques have greater chances of having better results on a wider field. In the biomedical field, supervised and semi-supervised techniques are focused on specific groups inside this field, such as semantic types for example. This happens because such systems rely on training data that most times does not exist and have to be developed to perform some tasks.

Knowledge-based approaches, specifically graph-based, do not require any human effort as they make use of existing ontologies or thesaurus, that already exist and are maintained, to perform disambiguation. Moreover, these systems can easily be upgraded by complementing or combining the knowledge sources. The effort associated with this task is practically none when compared to the effort of creating high quality training data for

supervised systems. Knowledge sources, such as UMLS, already cover a large portion of the biomedical domain. Also, it is well maintained and updated regularly with new data, thus enriching the knowledge. The richer the knowledge source the better results are.

Another means of improving results on such solutions is by simply fine-tuning or changing the scoring algorithms or, once again, combine multiple algorithms. This is a very simple task as the scoring phase of the disambiguation process is isolated from the remaining. A huge variety of graph-based algorithms to perform WSD were already presented. Navigli and Lapata [32], for example, studied several connectivity measures, such as *degree centrality*, *eigenvector centrality*, *key player problem*, *betweenness centrality*, and other global graph measures. Mihalcea [25] also presents some graph-based centrality algorithms such as *indegree*, *closeness*, *betweenness* and *PageRank.* From all these connectivity measures the simplest is *degree centrality,* in which a vertex is considered to be central if it has high degree. The vertex degree is given by the number of edges terminating in that vertex. Eigenvector centrality has two variants: *PageRank* and *Hypertext Induced Topic Selection (HITS)*. *PageRank* determines the relevance of a vertex by recursively assigning a score to each vertex. *HITS* determines two values for each node: the authority and the hub value. A vertex has high hub value if it points to many other vertices, while high authority represents a vertex that is linked by many good hubs. In an undirected graph, those values coincide.

Taking these facts into consideration, and with the vision of producing a powerful, scalable an easy to integrate tool, a graph based approach was chosen for this system.

3.2. Mission and Requirements

The main goal of this solution is to provide the biomedical community with a powerful tool that can help to improve the efficiency and consistency of information extraction. In order to do so, this system needs to provide an easy to integrate interface, thus being a secondary goal.

There are several annotation systems available nowadays for curators, that automatically map text to biomedical concepts – for example MetaMap [6], BeCAS [33], BioPortal annotator [34] – but most lack the WSD component. Some use the first occurrence found, others chose to return all possible occurrences or some other form of baseline for ambig-

uous cases. The importance of these tools in the biomedical domain is increasing due to the fast growing pace of the field and because they enable curators to be more efficient in the curation process. Providing such tools with a WSD component would significantly improve their performance.

Therefore, it is a requirement to develop a system that provides a simple application programming interface (API) in order to facilitate the integration in third-party systems, as well as enable developers to build their own applications that can interact with this.

This API must be able to receive data as input and generate disambiguated data in some standard format. For the system applicability to be wider, it must support various input data format, from RAW text to any specific predefined formatted data. The output format should be easy and lightweight so it can be easy to use and interpret by third-party software developers.

For this system, it is also important to take into consideration memory and processing limitations. The entire process of disambiguation is a complex task and, therefore, has high costs in terms of hardware requirements. Another key factor in this solution is its performance in terms of response time. For the system to be usable it must be able to provide results in a short span of time. For this reasons, all variables – memory, processing power and response time – must be evaluated together and an equilibrium point must be reached for the application to be efficient and usable.

3.3. System Architecture

In this section an architecture, based on the previous discussed requirements, is presented. In order to create an efficient software solution that answers all the requirements the technologies that support its core must be carefully chosen. Moreover, it is also important to optimize all algorithms and techniques to have a fast application and to save hardware resources.

Taking into consideration the great importance of an easy to integrate system, and following the trend in computer technology, the optimal approach to this problem is to provide a Web API. Such API is a programmatic interface, hosted on a web-server – most

commonly an HTTP web server – that implements a request-response message system. In this context REST[3] web-services stand out.

RESTful APIs provide easy and fast access to information along with simple integration with any development platform. Such APIs are typically defined with a base URL for the application endpoints (e.g. http://application.com/resources/), an Internet media type (e.g. JSON) for the data transferred and with standard HTTP methods (e.g. POST or GET). Table 3.1 presents the standard use for each HTTP method.

Table 3.1: HTTP methods and standard usage in a RESTful API.

HTTP Verb	http://application.com/resources/{id}
GET	Used to retrieve the representation of the requested resource in the defined media type.
PUT	Used to update data correspondent to the defined resource.
POST	Used to create a new instance of the resource, using the data sent.
DELETE	Used to delete information respective to the defined resource.

Using this type of web-service allows an easy and appropriate access to the data generated by the system. It is easy to note that the described HTTP methods are very general and are commonly used to interact with web based applications, however in the case of this system it is not necessary to implement all of the mentioned methods. Since the application will generate results based on data input from the client the only method necessary, and that will be implemented, is POST.

As mentioned, a correct Internet media type must be defined to transfer data between the server and the client. There are several media type standards registered and available[4]. In order to properly select the correct media types to allow in this system, it is important to analyze the input and output data. As mentioned before, the application must support various input data types and will output formatted data to the client. Con-

[3] REST – Representational state transfer

[4] Official registry of media types: https://www.iana.org/assignments/media-types/media-types.xhtml

sidering that such data may be RAW text or any commonly used format in the biomedical field, which includes XML formats and CSV based formats, the selected media types are **`text/plain`**[5], **`text/xml`**[6] and **`application/json`**[7].

Lastly, a programming language must be selected to develop this solution. In order to assure cross compatibility and to make sure that the application can be deployed and integrated in the most variable system setups, a multi-platform compatible language must be chosen. For this purpose, Java was chosen. Not only is it cross compatible but also it allows the development of object-oriented, concurrent solutions.

This solution relies on a knowledge source, as discussed before, that must be stored and must be accessible to the application. For this purpose a graph database could be used, however a study of the available database systems concluded that relational databases perform better than graph databases (see figure 2). For this reason, a MySQL RDBMS[8] was used, in order to store data used by the system – in this case the UMLS database. The access to data stored by this database is granted using the JDBC[9] technology. This is a Java API that connects and allows the execution of queries to a database. It is a widely used technology as it can connect to most database types, either locally or remote, granting the possibility to work with databases available from third-party systems.

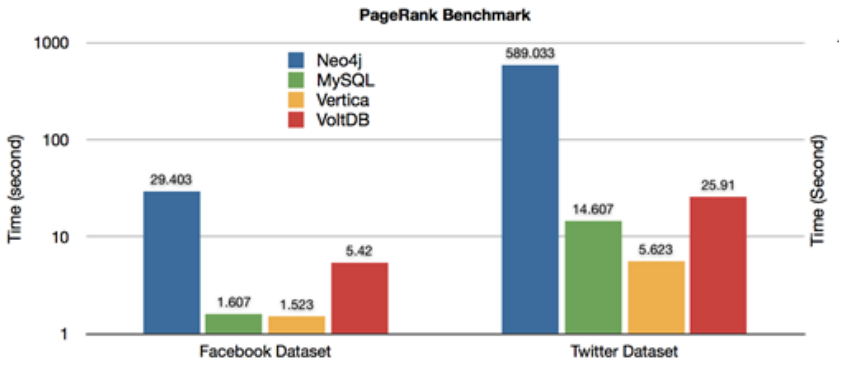

Figure 2: PageRank on different systems (Adapted from ISTC for Big Data Blog[10])

[5] Plain textual data
[6] eXtensible Markup Language data
[7] JavaScript Object Notation data
[8] Relational Database Management System
[9] Java Database Connectivity
[10] http://istc-bigdata.org/index.php/benchmarking-graph-databases/

Finally, since this proposal makes use of graphs to produce results, an efficient algorithm to process graphs must be developed. For this purpose, JUNG was used to help manage and model graphs. JUNG[11] is the Java Universal Network/Graph Framework, which is an extensive library used to model, analyze and/or visualize data in the form of graphs or networks. This library allows the creation of various types of graphs, including directed and undirected graphs, multi-modal graphs, hypergraphs and more. It also allows the use of custom data types for vertices and edges, which is important for this solution since it allows modeling vertices and edges to match the UMLS data structure.

3.4. Implementation

Bringing together all the technology selected, and discussed in the last section, and building an efficient system, in terms of speed, memory usage, and processing effort, is a huge challenge. In this section the management of hardware resources, as well as software implementation specific considerations are further explained.

Regarding hardware resources it is important to take into consideration disk usage, memory usage and CPU usage. Disk usage is related to the database, the bigger database the bigger disk usage. Since the UMLS database is very complete and extensive, and given the fact that this system makes use of relations between concepts, all data that is not used by the solution can be deleted. For this purpose, in order to use the least possible disk only the tables used from UMLS were imported into the database used by the application. Since MRREL is the table that contains all relations between concepts that are needed to create the graphs, all other tables were not imported.

Despite the fact that only one table is present in this database, it is important to note that this table occupies approximately 10GB of disk space, containing 38,935,715 relations. It is perceptible that searching relations between concepts in such a large dataset requires big efforts from CPU and disk. For this reason, a way to optimize search queries had to be studied. The concept of covered index was applied to boost the performance of the system.

[11] http://jung.sourceforge.net/

A covered index (Figure 3) is a specific type of index where the index itself contains all data fields used in the statement. If a normal index were used, when performing a query the system could find the results quickly in the index but afterwards would have to address the database in order to retrieve the remaining needed fields. For example, assuming that the system needs to find all concepts in relations starting at concept A. It would search the index and next would have to retrieve the opposite concept of the relation from the database, hence having to address the table itself. Using a covered index brings the ability to find all concepts related to concept A just by searching the index, which is expected to be much smaller than the table, providing significant speedups [35].

```
ALTER TABLE MRREL ADD INDEX X MRREL CUI1 (CUI1, CUI2)
```

Figure 3: Example statement for creating a covered index in MySQL

Graph operations are computational heavy and time expensive. In order to speed up these processes it is optimal to have all data to be analyzed in memory, as memory operations are many times faster than operations that require disk access. However as it is easy to understand, looking at the size of the dataset, it is impossible to keep all data in memory. For this reason an algorithm was developed (see section 3.5) to create an in memory graph, referred to as the general context graph, given a set of starting vertices. In this case, the starting point is given by the set of concepts present in the instance to be disambiguated.

Because the purposed solution addresses the problem of WSD and not the annotation process, and because it is a requirement that the system must be able to process free text, i.e., text that has no information about concepts, a solution to pre-process this data is needed. To address this question BeCAS [33], a system developed at the University of Aveiro Bioinformatics group by Nunes et al., is used. BeCAS provides a web API for biomedical concept identification. It takes as input free text and generates text annotated with biomedical concepts in one of the chosen formats from the set of available formats –

JSON, XML, A1[12] or CoNLL[13]. An interface able to communicate with BeCAS API was implemented in the proposed solution in order for it to be able to process free text as input.

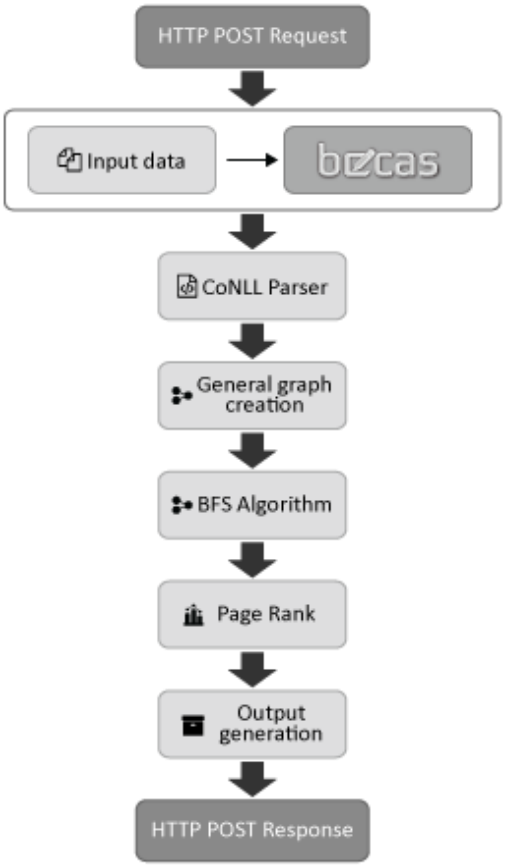

Figure 4: General processing pipeline of the system

The purposed solution is divided in 4 general phases (Figure 4): Parsing; Graph generation; Scoring; Output generation.

For the first part, a parser, that retrieves the list of concepts present in each sentence of the input text, was developed. This list contains the CUIs associated with each of the terms identified in the sentence where, in case of ambiguity, a term is associated to a list of the respective candidate CUIs. Each sentence is, therefore, represented by a list of

[12] A1 is a format used in text-mining research tasks

[13] CoNLL format includes sentence splitting, tokenization, lemmatization, part-of-speech (POS) tagging and chunking, in addition to concept identification.

concept lists, where a concept list containing a single element represents an unambiguous term.

In the proposed system, the parser developed takes as input CoNLL (Figure 5) format, meaning that in cases where the input is RAW text, BeCAS is used to generate the needed formatted data. Moreover, because of a modular and scalable development approach it is possible to easily develop and implement parsers for other formats which can be done as future work if necessary.

```
1  Immunohistochemistry  Immunohistochemistry  B-NP  NN   0    _
   _    _    _
2  using use  B-VP  VBG  0    _    _    _    _
3  an   an   B-NP  DT   0    _    _    _    _
4  antispirochete  antispirochete  I-NP  JJ   0    _    _    _
   _
5  (    (    O    (    0    _    _    _    _
6  T    T    B-NP  NN   0    _    _    _    _
7  .    .    I-NP  FW   0    _    _    _    _
8  pallidum  pallidum  I-NP  FW   UMLS:C0017651:T023:ANAT    _
   _    _    _
9  and  and  O    CC   0    _    _    _    _
10 Borrelia  Borrelia  I-NP  NN   UMLS:C0024198:T047:DISO|
    UMLS: C0006033:T007:LIVB  _    _    _    _
11 )    )    O    )    0    _    _    _    _
12 antibody  antibody  B-NP  NN   UMLS:C1621287:T044:PROC|
    UMLS:C0021027:T129:CHEM|UMLS:C0003241:T129:CHEM|
    UMLS:C0021027:T121:CHEM   _    _    _    _
13 was  be   B-VP  VBD  0    _    _    _    _
14 performed  perform   I-VP  VBN  0    _    _    _    _
15 retrospectively  retrospectively  B-ADVP    RB   0    _    _
   _    _
16 .    .    O    .    0    _    _    _    _
```

Figure 5: Example instance of the ambiguous term “Borrelia” in CoNLL format

The second phase is the most computational heavy, as it its responsible for graph generation and processing. The lists generated by the parser are iterated in this step and, for sentences containing at least one ambiguous term (i.e. a concept list containing more than one element) a general context graph is created in memory, using the algorithm described in Section 3.5. This graph contains all possible concept identifiers associated to each term in the sentence, all UMLS concepts that have relations (in the knowledge-base)

with those concepts, as well as all relations between those concepts. Most of these relations and concepts are unnecessary for the scoring phase of the solution, however in this phase they are crucial to find all possible connections between the target concepts and the context concepts.

Thereafter, it is needed to extract the relevant concepts and relations from the context graph. For this purpose, a breadth first search (BFS) algorithm is implemented to retrieve all possible paths between concepts present in the sentence to be disambiguated. In this step, each concept list is iterated in two different cycles in order to have the *to* and *from* vertices to use in the BFS algorithm, ensuring in this way that all paths between each of the concepts in the sentence are retrieved. All paths found by this algorithms are then used to create a final graph that contains only relevant vertices and edges, thus ending the second phase.

Because of the density of the MRREL table a limit for the graph size had to be set to preserve memory. In order to limit its dimension a maximum depth parameter is defined. For the same reason a maximum hops parameter for the BFS algorithm was also set, establishing a maximum number of concepts between the start and end nodes. Both parameters can be adjusted to different values to fine tune results.

The third phase of the system is responsible for scoring the resulting graph. For this process several scoring algorithms exist, as discussed before (see section 3.1). Thanks to the modularity of the system, these scoring algorithms can easily be replaced or fine-tuned in order to improve or combine results. Nevertheless, for the purposed solution, the scoring algorithm used was PageRank. Not only this algorithm reveals to have one of the top performances [32] but it also enables the developer to increase or decrease the score of the graph vertices based on a given criteria, thus being able to get more accurate results.

Lastly, the output generation phase is responsible for compiling all data from all sentences and format it in a specific format that is then sent as response to the original POST request issued. The output format is a JSON object, containing the sentences and respective identified CUIs for the identified terms. In case of ambiguities, the CUI of the ambiguous term with highest page rank score is returned.

Analyzing all the steps needed to disambiguate a single instance, it is easy to deduce that this process will have high CPU and memory usage. In order for this system to be able to run on most servers, it is possible to configure the maximum number of instances running at the same time, i.e., the system can have two or more simultaneous processes running if the server has the hardware to support it.

The resulting developed system is available at http://bioinformatics.ua.pt/biowsd/ for free usage. Figure 6 represents an example input object for the POST method of the API, and Figure 7 shows the resulting output object.

```
{
  "input_type":
    "RAW",
  "input":
     "Despite great efforts devoted to clarifying the
     localization of proliferative activity in the adrenal
     cortex, the agents that stimulate proliferation remain
     controversial, and the nature of the stem cells from which
     cortical cells differentiate is incompletely
     understood."
}
```

Figure 6: Example POST input object

```
{
 "output": [
  {
   "text": "Despite great efforts devoted to clarifying the
localization of proliferative activity in the adrenal
cortex, the agents that stimulate proliferation remain
controversial, and the nature of the stem cells from which
cortical cells differentiate is incompletely understood.",
   "line": 1,
   "concepts": [
    {
     "text": "adrenal",
     "cui": "C0001613",
     "end": 101,
     "start": 94
    },
    {
     "text": "cortex",
     "cui": "C0001613",
     "end": 108,
     "start": 102
    },
    {
     "text": "cells",
     "cui": "C0007634",
     "end": 206,
     "start": 201
    },
    {
     "text": "cortical",
     "cui": "C0001613",
     "end": 226,
     "start": 218
    },
    {
     "text": "cells",
     "cui": "C0007634",
     "end": 232,
     "start": 227
    }
   ]
  }
 ]
}
```


Figure 7: Example API output object

3.5. Algorithms

This section describes the algorithms used in the developed solution. The algorithms to be described are the graph generation algorithm and the ranking algorithm. The first makes use of queries to the database to generate a graph based on the input instance, the latter is a PageRank based algorithm.

The graph generation algorithm (Algorithm 3.1) can be seen as a two steps algorithm, where the first step is to retrieve the general graph that includes the set of admissible meanings present in the instance to be disambiguated. The second step is to limit this graph by extracting the paths between each meaning, in order to remove unnecessary information from it.

The input of this algorithm is a list that contains a set of possible meaning for each annotated concept. At first, this list is used to retrieve all related concepts which are then analyzed, added to graph – along with respective relations – and compiled into a new list of unexplored concepts. Those will be used in the next iteration to retrieve the list of related concepts. This process is repeated for a maximum number of iteration or until no more unexplored concepts are found. Afterwards, using BFS algorithm, all paths between the input set of meanings are retrieved and added to a new graph – Figure 8 illustrates a sample graph resulting from this algorithm. The first graph is then deleted and the latter will be used to perform scoring techniques and retrieve results.

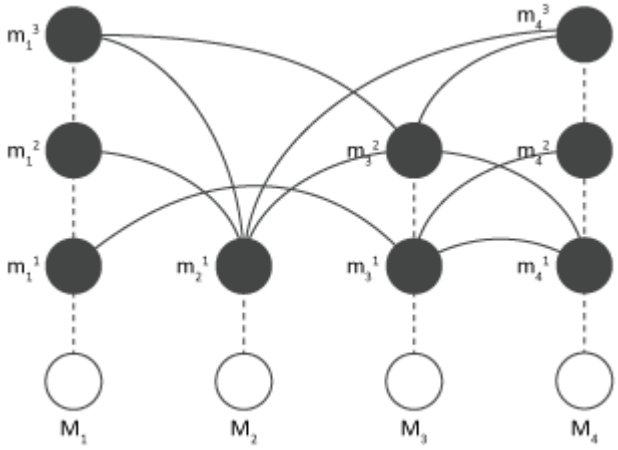

Figure 8: Sample graph created by algorithm 3.1 (adapted from [25])

Algorithm 3.1: Graph generation algorithm

Input: Admissible meanings for each word $M_i = \{m_i^t \mid t = 1 .. N_i\}$, $i = 1 .. N$

Output: Sequence of chosen meanings $M = \{m_i \mid i = 1 .. N\}$, with meaning mi corresponding to the highest scored meaning from the input admissible meanings.

Build general context graph G

```
 1:   search_input ← M
 2:   for MaxDist to 0 do
 3:     discovered ← NewSet()
 4:     relations ← QueryDatabase(M)
 5:     for i = 1 to N_relations do
 6:       AddEdge(G, relations_cui1, relations_cui2, relations_rel)
 7:       if relations_cui1 not in M then
 8:         Add(discovered, relations_cui1)
 9:       end if
10:       if relations_cui2 not in M then
11:         Add(discovered, relations_cui2)
12:       end if
13:     end for
14:     search_input ← discovered
15:   end for
```

Build context graph C

```
 1:   for i = 1 to N do
 2:     for j = i + 1 to N do
 3:       for t = 1 to N_i do
 4:         for s = 1 to N_j do
 5:           Paths ← BFS(G, m_i^t, m_j^s)
 6:           for k = 1 to N_p do
 7:             AddEdge(C, Paths_cui1, Paths_cui2, Paths_rel)
 8:           end for
 9:         end for
10:       end for
11:     end for
12:   end for
```

On the other hand, algorithm 3.2 is the scoring technique used in this solution. As discussed before, there exist several choices for scoring algorithms, however the chosen was PageRank. For this system, a Personalized PageRank algorithm was implemented [26]. Initially all graph vertices have the same level of importance, meaning that each vertex has the same initial PageRank value, in this case the value 1 (one). Some techniques may be developed to fine-tune this algorithm by giving greater importance to specific

vertices based on a given criteria. This personalization may improve the success rate of the application.

This scoring technique is based on a random-walk model, where a random surfer takes random steps on the graph. Such walk can be modeled as a Markov process, i.e., the decision on what edge to follow depends on the current vertex only. The PageRank score of a vertex means the probability that the random surfer is found on that vertex, assuming an infinite walk. After a certain number of walks, the probabilities converge to a stationary distribution.

Being G the graph with vertices (V_1,…,V_n). For a given vertex V_i, *Out*(V_i) is the set of vertices outgoing from it and *OutDegree*(V_i) is its out-degree (i.e., count of outgoing vertices). The PageRank of V_i is given by:

$$PR(V_i) = (1 - d) + d \times \sum_{V_j \in Out(V_i)} \frac{PR(V_j)}{OutDegree(V_j)} \quad (1)$$

Where *d* is the *damping factor*, a scalar between 0 and 1, typically set to 0.85 by default.

3.6. Summary

In this chapter an in-depth overview of the solution proposed was discussed. Firstly the reasons that support choosing a graph-based approach to the WSD problem are explained, emphasizing that such solutions are less expensive in terms of human effort than supervised approaches. Then the goal and requirements of the application are presented,

Algorithm 3.2: PageRank scoring algorithm

Input: Context graph G.

Output: Graph G with vertex scored with PageRank probabilities.

Score vertices in graph G

```
1:   for Vi in Vertices(G) do
2:     for Vj in Out(Vi) do
3:       successors_sum ← successors_sum + PR(Vj) / OutDegree(Vj)
4:     end for
5:     PR(Vi) ← (1 - d) + d * successors_sum
6:   end for
```

giving great importance to the capability of building an easy to integrate system in order to provide the biomedical community with a powerful solution. For this purpose, following the nowadays trends in computer science, it was concluded that a web application programming interface should be developed to accomplish an easy-to-use and easy-to-integrate system.

Afterwards, in section 3.3 and 3.4, a more technical insight on the proposal is given. Firstly explaining the reasons of choosing a Web API as solution and presenting the general working architecture of such systems. Focusing on the developed solution, it was explained that a simplified API would be developed, implementing only POST methods and supporting 3 (three) different media type for the data transferred. Then the implementation of the proposal is further detailed and explained, presenting a system pipeline (Figure 4) and explaining the specific algorithms used (Algorithm 3.1 and Algorithm 3.2).

Chapter 4

Experiments and Results

In this chapter the experiments and results of the proposed solution are addressed. Firstly the dataset used to validate the system is analyzed. Next some general considerations about the experiments are discussed and lastly the experiments are described, including their methodology and results followed by a discussion. The goal of the experiments is to analyze the success rate of the solution, comparing with state-of-the-art systems in order to evaluate it.

4.1. Experimental Dataset

This section presents a detailed description of the dataset used to evaluate the proposed system, specifically the National Library of Medicine (NLM) Word Sense Disambiguation Test Collection. This dataset was first developed in 1999 [5] for the purpose of supporting researchers in the task of testing their disambiguation solutions. More recently, NLM released a second version of the dataset, called MSH WSD Test Collection [36].

The first collection contained 50 ambiguous words corresponding to 50 ambiguous concepts from the 1999 UMLS. This dataset was built by compiling about 5,000 citations from 1998 MEDLINE, representing 5,000 ambiguity cases. Each ambiguous instance was resolved by hand by a group of curators that examined each of the citations.

The latter collection, MSH WSD Test Collection, was created automatically by extraction instances of ambiguous terms from MEDLINE and using MeSH indexing of MEDLINE as a resource. The resulting dataset contains 203 ambiguous words present in the 2009AB UMLS, annotated on more than 37 thousand 2010 MEDLINE citations.

For the purpose of testing the proposed solution, the most recent version of the collection was used. This data set is much larger and richer than the previous. Each of the 203 words is associated with two or more possible CUIs (Concept Unique Identifiers) from the UMLS. For each possible CUI, a set of citations containing the given word was compiled. The data is available in plain text and follows the Attribute-Relation File Format (ARFF[14]). Figure 9 shows a typical instance of an ambiguous term. The data fields are separated by a comma and represent the PubMed ID, the citation (title and abstract) and the associated meaning, in this order.

```
@RELATION C0024198_C0006033

@ATTRIBUTE PMID integer
@ATTRIBUTE citation string
@ATTRIBUTE class {M1, M2}

@DATA
20509909,"Prevalence of <e>Borrelia</e> burgdorferi sensu lato in
rodents from Gansu, northwestern China.BACKGROUND: Lyme disease
is a multi-organ infection disease caused by Borrelia burgdorferi
sensu lato. Lyme disease was first documented in north-east China
in 1986. Since then more than 20 provinces in China were
confirmed the existence of nature foci of Lyme disease. In the
present study, a molecular epidemiological survey was conducted
to investigate the presence of Borrelia burgdorferi sensu lato in
rodents from Gansu Province for the first time. RESULT: A total
of 140 rodents of 7 species were examined for Borrelia
burgdorferi sensu lato. by nested-PCR and culture isolation. The
overall infection rate was 22.86%. Two rodent species most
frequently trapped were responsible for all positive. 3 strains
were isolated from Apodemus agrarius, which belonged to B.
garinii, 1 strain isolated from Rattus losea was identified as B.
afzelii. CONCLUSION: The study firstly showed the role of rodents
in maintaining the pathogen of Lyme disease in the environment
from Gansu Province and there existed at least two genotypes of
Lyme disease spirochaetes in rodents.",M1
```

Figure 9: An instance of the ambiguous term *Borrelia* in the MSH WSD Test Collection

[14] http://www.cs.waikato.ac.nz/ml/weka/arff.html

In each instance, the target word is denoted by the *e* tag (e.g. *<e>Borrelia</e>*). There are a total of 37,888 instances spread over 423 different possible meanings (CUIs) for the 203 ambiguous words.

Table A.1 shows the distribution of the possible senses of each word in the dataset. M1 through M5 are different sense labels as defined in the ARFF file of each word. The last column presents the total number of instances for each word. In table A.2 the possible CUIs for each meaning is displayed. This dataset also covers a broad range of semantic types, precisely 82 different types. Table A.3 shows the semantic types present in this collection, along with the frequency in number of concepts.

4.2. Experiments General Considerations

In order to feed the above descripted data into the proposed software solution, it was required to pre-process the data as well as adapt the application to remove the overhead of data transfer via HTTP protocol.

Firstly, regarding data pre-processing, it was necessary to retrieve all citations from the ARFF files into separate files, each containing an individual instance. For this purpose, a bash script was developed that iterates over each ARFF file, creates a directory with its name and extracts each instance into individual files. Each file has the citation (title and abstract) as content and the PMID as file name. For the system to be able to validate the results, i.e., to verify if the predicted meaning matches the meaning provided by the dataset, it was also necessary to create a mapper file that would map each PMID to the correct CUI provided by the collection.

After having the individual instances separated by files, because the only provided data with the instances is the identification of the target word, each citation had to be annotated to identify mentions of biomedical concepts. For this purpose, and to remove the overhead of data transfer via HTTP, Neji [37] was used to annotate all instances. Neji is a modular tool used in the process of annotating biomedical text while allowing the user to define various customization parameters. It integrates a powerful command line interface (CLI) tool, that provides a complete set of features, such as: annotate using dictionaries and/or machine-learning models with respective normalization dictionaries; choosing from a set of various input and output formats; number of threads customiza-

tion; etc. All instance files were annotated using this tool, providing it with the necessary dictionaries, selecting the most appropriate output format (in this case CoNLL) and setting an optimal number of threads to run in order to speed up Neji as much as possible. Because some semantic types covered by the dataset are not considered important, a subset of the collection was not covered (see table A.4 for uncovered data types). Thus a total of 74 out of 203 concepts were removed from the results considerations.

Table 4.1: Terms excluded from results

Terms		
AA	Digestive	Nursing
ADA	DON	OH
ADP	drinking	ORI
ANA	eCG	PAC
BR	Eels	Pharmaceutical
Brucella abortus	EGG	pI
BSA	EM	Platelet
BSE	EMS	POL
CAD	ERP	Potassium
Cardiac pacemaker	Erythrocytes	PR
Cell	Exercises	Projection
Cement	Fish	PVC
CH	FTC	Radiation
Cholera	HR	RBC
CI	Hybridization	Sodium
Cilia	IA	SPR
CIS	INDO	STEM
CNS	IP	TAT
Coffee	JP	Tax
Compliance	LABOR	TEM
cRNA	Language	TLC
Crown	Laryngeal	tomography
dC	MRS	US
DE	NM	veterinary
DI	Nurse	

After having the output of Neji, i.e. annotated instances with concepts identified, it was necessary to use this data as input of the proposed system. Again, because using HTTP transfers is unnecessary the application was adjusted to accept input from files stored locally in the same machine. Using a CLI interface it is possible to configure the path to the data files as well as the path to a mapper file used to validate the results obtained. Another change made in the software was introducing the restriction to prevent disambiguation of the full citation, i.e., the only disambiguation performed by the system in this mode is regarding the target word. At the end of each disambiguation iteration, the obtained result is then compared with the result retrieved from the mapper file and an output CSV[15] file is generated that stores the PMID of the citation, the term chosen by the system and the term in the mapper file, which is the correct meaning for the given citation. A different CSV file is generated for each one of the 203 ambiguous, as the data files respecting each word are stored in different folders to preserve the concepts separation and allow an easier analysis.

In the following sections the results from the various experiments performed are presented and discussed. A score is given for each term considered and a total average of the results is presented. The scores will be presented in the form of percentage, comparing the number of correct predictions versus the total number of instances for each term (accuracy).

4.3. UMLS 2014 Experiment

For this experiment the UMLS 2014AB version was used. This experiment will be used as baseline for the remaining experiments of this dissertation. All data from MSH WSD corpus was annotated using the above mentioned methods and all concepts were disambiguated having the version 2014AB of UMLS as knowledge source. Despite the fact that all terms in the dataset were disambiguated using the proposed system, only results for the 129 considered terms are presented and discussed in detail. Table 4.2 shows the overall obtained score using this approach. Comparing to state-of-the-art performance (Section 2.3) this results are closely behind. One can conclude that with some improvements better results can easily be obtained.

[15] Comma-separated values

Table 4.2: Overall result of UMLS experiment

Experiment	Number of Instances	Correct Predictions	Overall Score
UMLS 2014	24301	14137	0,5817

Table 4.3 presents a list of the five best and worst obtained results (refer to table B.1 for a detailed overview of the results). In this table some satisfactory results are shown, with a total of 12 terms with a prediction accuracy above 85%. These terms show very good results thanks to a rich and well balanced amount of relations starting from the ambiguous terms. Also a very rich context for each ambiguous instance is very important to accomplish the best results: for example, the terms *CLS* and *PCD* have an average of 21 concepts per phrase.

On the other hand, some terms show very low results with 4 terms having an accuracy of less than 30%. In these cases, the amount of relations starting at the ambiguous concepts is under the average, resulting in a poor network unable to correctly disambiguate the terms. Furthermore, in some cases the number of outgoing edges from the possible senses is unbalanced causing the system to have a uneven network. For example, the term *Synapsis* has 216 edges from C0039062 (M1) and only 45 from C0598501 (M2), causing the system to appoint the first meaning as the correct more easily. Because the sense distribution of such term is not even – M1 has 35 instances present in corpus, whereas M2 has 99 – the final result is, as shown in table 4.3, a total of 38 correctly predicted instances. Lastly, a small number of concepts per phrase was noted for most of the terms with low accuracy result, with an average of 9 concepts per phrase for term *Hemlock.*

Table 4.3: Best (left) and worst (right) 5 results for UMLS

Term	Total Instances	Correct Predictions	Score	Term	Total Instances	Correct Predictions	Score
CLS	34	34	1,0000	**RA**	297	94	0,3165
Follicle	198	190	0,9596	**Ca**	396	118	0,2980
PCD	198	188	0,9495	**Synapsis**	134	38	0,2836
HPS	178	163	0,9157	**Hemlock**	77	20	0,2597
Glycoside	198	180	0,9091	**Lawsonia**	115	15	0,1304

In summary, the overall score can and must be improved despite the fact that 58% correct predictions is satisfactory in such a large dataset with a wide range of semantic types. However, there exist a few situations where the number of correctly disambiguated instances is much lower than the average. Some work must be done in order to fill this gap and improve score on this cases, thus greatly improving overall results.

4.4. Experiment using Most Frequent Sense (MFS)

In order to improve the results obtained in the previous experiment, a new approach to scoring the vertices was taken. Properly scoring the graph is crucial to obtain better results. For this purpose, a study of the most common senses of each ambiguous term was conducted. Such was achieved using frequency counts from MEDLINE 2015 Baseline. Because such frequency counts are related to MeSH terms, an algorithm was developed to map each CUI of the ambiguous terms to the respective MeSH Unique ID, thus creating a file containing CUIs from MSH WSD corpus and their respective frequency count.

Thanks to the modular approach of the proposed application, the disambiguation workflow was not affected being all the changes made in the scoring algorithm, in order to use this new information. The scoring basics of the algorithm are the same, giving a vertex score of 1 (one) to each vertex and adding a maximum increment of 0.5 in each of the vertices on the ambiguous concept. This increment is calculated relatively to the frequency count, for example if *Meaning A* has a frequency count of 1000 and *Meaning B* has a count of 600, the first will receive a boost of 0.5, while the latter sums 0.3 to the initial score.

Table 4.4: Comparison of overall results of experiments

Experiment	Number of Instances	Correct Predictions	Overall Score
UMLS 2014	24301	14137	0,5817
MFS	24301	14402	0,5927

Table 4.4 compares the overall accuracy of the experiment discussed in section 4.3 and the one presented in this section. It proves that using an improved scoring technique

will affect positively the overall result of the proposed system. Despite the fact that the overall accuracy merely improved by 1%, observing table 4.5 it is noticeable that the worst results in this experiment suffer a great improvement when compared to the previous. In this scenario, the term *Lawsonia* doubled the number of correct predictions achieving a final accuracy of 28.7%. Also looking that the detailed table (table B.2), the term *Hemlock* doubled the accuracy obtained in the previous experiment from 25.97% to 51.94%.

On the other hand, some terms suffered a drop in the results. For example, term *HPS* declined from 91.57% to 88.2%. Analyzing the complete results, term *Glycoside* suffered the largest decrease of accuracy from 90.9% to 79.3%.

Table 4.5: Best (left) and worst (right) 5 results for MFS

Term	Total Instances	Correct Predictions	Score	Term	Total Instances	Correct Predictions	Score
CLS	34	33	0,9706	**RA**	297	101	0,3401
Follicle	198	191	0,9646	**lens**	295	98	0,3322
Follicles	198	190	0,9596	**Ca**	396	118	0,2980
PCD	198	188	0,9495	**Lawsonia**	115	33	0,2870
TRF	179	163	0,9106	**Synapsis**	134	36	0,2687

In conclusion, this approach proves to achieve better overall results than using the UMLS 2014 without any extra information. The improvements in the 5 worst results are notable, as well as overall, having terms with an increase of accuracy in the order of 15% to 25%. These improvements are more noticeable in terms that proved, in the previous experiment, to have a poor context giving a boost to the most commonly used meaning. Despite the fact that some terms suffered a decrease of accuracy, these cases are less significant than the improvements obtained (maximum decrease is 11.62% for term *Glycoside*). Such decreases are more common in terms that have a balanced distribution of instances for each meaning, because boosting the most frequent sense will lead to errors in some predictions of the instances that refer to less frequent terms.

4.4.1. Acronym Disambiguation

Analyzing the terms present in this corpus, it is observable that some of the considered terms are acronyms (60 acronyms in 129 total terms). In most cases of acronym usage, the long form of the word is also used in text at least once. Therefore, an algorithm was developed to identify the occurrence of the long-form of the acronyms in the instances considered.

Once more, using the modular structure of the system no major changes were needed. An algorithm responsible for identifying the acronym expansion in text and boosting the respective vertex in graph was implemented. This new module is used after the scoring phase, therefore the scoring basics are as described in the previous section.

In order to identify the acronym long-form in text, this module uses the list of concepts obtained from the phrase to disambiguate and removes the CUIs of the concepts related to the target term (i.e. the acronym) from this list. Afterwards, it uses the CUIs associated with the acronym and searches for the occurrence of each in the remaining list of concepts. If an unambiguous term matches any of the acronym candidate CUIs then it should be the long-form of the acronym, therefore boosting the respective vertex in the graph.

Table 4.6: Comparison of MFS with acronym disambiguation with previous experiments

Experiment	Number of Instances	Correct Predictions	Overall Score
UMLS 2014	24301	14137	0,5817
MFS	24301	14402	0,5927
MFS – Acronym Disambiguation	24301	14627	0,6019

Despite the fact that the global outcome of this experiment was positive (approximately 0.9% increase in accuracy), a larger improvement was expected. This was not the case, mainly because the disambiguation scope of the system is based on sentences. In some cases, the long-form of the acronym would appear on a different sentence from the one being disambiguated, for this reason expanding the scope to the paragraph or even

the whole document could be beneficial. However, it is important to understand that doing so has costs in terms of efficiency.

Nevertheless, analyzing table 4.7 two acronyms, *CCD* and *BPD*, appear in the top 5 that were not present in the top 5 of the previous experiments. These two terms had an increase in score from around 85% to about 99%, ending with only one and two wrong predictions, respectively. This represents a good improvement in the disambiguation results of acronyms. Moreover, looking at the detailed table (table B.3) many acronyms are identified where the correct number of predictions did not change. As mentioned above, this mainly happens because of the scope of disambiguation.

Table 4.7: Best (left) and worst (right) 5 results for MFS – Acronym Disambiguation

Term	Total Instances	Correct Predictions	Score	Term	Total Instances	Correct Predictions	Score
CCD	141	140	0,9929	**RA**	297	101	0,3401
BPD	198	196	0,9899	**lens**	295	98	0,3322
CLS	34	33	0,9706	**Ca**	396	118	0,2980
Follicle	198	191	0,9646	**Lawsonia**	115	33	0,2870
Follicles	198	190	0,9596	**Synapsis**	134	36	0,2687

Concluding, this new module has proven to be useful to correctly disambiguate acronyms. The results can be further improved by enlarging the scope of disambiguation to the paragraph or document, however such change must take into consideration the application efficiency as well as system hardware limitations.

4.5. MeSH Terms Experiment

In the previous sections, techniques and approaches are discussed and presented to improve the baseline accuracy scores. Such improvements, despite being small, prove that the proposed knowledge based solution has a lot of potential. For this reason, a new experiment was conducted using another technique trying to improve the previously demonstrated results.

Focusing on the scoring phase of the disambiguation process has proved to be effective in the enhancement of the results. Therefore, a new algorithm for scoring the initial graph was developed. In this case, the system would search in PubMed the 100 most similar documents to the citation being disambiguated and extract the MeSH terms present in them. PubMed comprises over 24 million biomedical citations from MEDLINE, online books and life science journals. Each citation contains a list of MeSH terms identified within the citation. This list is retrieved from each of the top 100 most similar documents and compiled together. From the compilation, a score for each term is calculated and summed in the terms present in the context graph, boosting the most probable meaning. Because the extracted list contains more than the term to be disambiguated, the remaining terms are also boosted, this way giving greater importance to such terms in order to improve the final outcome of the scoring algorithm.

This experiment takes the same baseline as the discussed in section 4.3, being the scoring basis the same, i.e. each vertex starts with the same score. Then the above described algorithm boosts some of these vertices with a score from 0 (zero) to a maximum of 1 (one).

Table 4.8: Comparison of MeSH terms experiment with previous experiments

Experiment	Number of Instances	Correct Predictions	Overall Score
UMLS 2014	24301	14137	0,5817
MFS	24301	14402	0,5927
MFS + Acronym expansion	24301	14627	0,6019
MeSH	24301	15198	0,6254

In table 4.8 a constant improvement in results is displayed, with the MeSH experiment achieving the best scores. This experiment shows an increase of over 4% from the baseline. Once again it is noticeable that improving the scoring techniques results in better accuracy results. Analyzing the fully detailed table (table B.4), it is observable that only few results are negatively affected, being 7.6% the biggest decline in accuracy. On the other hand, many results were improved with a maximum increase in accuracy of 34.73% for term *TNC.* When comparing to the baseline experiment, table 4.9 shows a good im-

provement in accuracy of the top 5 best results which has a great impact in the final overall score of the system. On the other hand, the top 5 worst results show small improvements, showing a term that decreased accuracy (*Ca*) and a term that was not affected (*Lawsonia*). Such outcomes for these terms were expected due to the poor context in their citations, as well as the small amount of relations between the ambiguous concepts and the surrounding ones.

Table 4.9: Best (left) and worst (right) 5 results for MeSH

Term	Total Instances	Correct Predictions	Score	Term	Total Instances	Correct Predictions	Score
CLS	34	34	1,0000	**Cold**	259	92	0,3552
BPD	198	193	0,9747	**lens**	295	100	0,3390
OCD	198	193	0,9747	**Synapsis**	134	44	0,3284
HPS	178	171	0,9607	**Ca**	396	112	0,2828
PCB	127	122	0,9606	**Lawsonia**	115	15	0,1304

In summary, a great overall improvement is noticed, with 4 terms increasing more than 30% when compared to the baseline. Despite the efforts to increase the final overall result, the terms with lowest accuracy in the baseline experiment did not improve much. To overcome these issues represents a big challenge that can be further explored in future work.

4.5.1. Acronym Disambiguation

Taking into consideration the ideas discussed previously in section 4.4.1, it was considered that using the acronym disambiguation algorithm in this experiment can add value to the purposed approach. For that reason, the exact same principle as before was used to identify the acronyms long-form and boost the respective graph vertices.

The results are as shown below (table 4.10), an increase in accuracy of about 0.8%. In this experiment, the final result was expected because of the conclusions from the previous acronym disambiguation experiment. The improvement is slightly lower than the previous achieved increase which was expectable because PubMed contained MeSH terms

associated with the long-form acronym, thus increasing most acronyms accuracy (see table 4.9, where best 5 results are acronyms).

Table 4.10: Comparison of MeSH terms with acronym disambiguation experiment

Experiment	Number of Instances	Correct Predictions	Overall Score
UMLS 2014	24301	14137	0,5817
MFS	24301	14402	0,5927
MFS – Acronym Disambiguation	24301	14627	0,6019
MeSH	24301	15198	0,6254
MesH – Acronym Disambiguation	24301	15384	0,6331

Table 4.11 presents very good results for acronyms, showing two cases of 100% accuracy and other three over 96%. Analyzing the fully detailed table (table B.5), 25 of the 60 acronyms have a final score over 75% which is a very satisfactory result. In the 5 worst results, no terms were affected as none of it are acronyms.

Table 4.11: Best (left) and worst (right) 5 results for MeSH – Acronym Disambiguation

Term	Total Instances	Correct Predictions	Score	Term	Total Instances	Correct Predictions	Score
BPD	198	198	1,0000	**Cold**	259	92	0,3552
CLS	34	34	1,0000	**lens**	295	100	0,3390
CCD	141	140	0,9929	**Synapsis**	134	44	0,3284
OCD	198	193	0,9747	**Ca**	396	112	0,2828
HPS	178	171	0,9607	**Lawsonia**	115	15	0,1304

In summary, the combination of MeSH terms with the acronym disambiguation technique results in satisfactory overall results achieving a total score over 63%. With 17 terms achieving an accuracy over 90%, the main drawback of the system is the terms where the graph network is poor and unbalanced resulting in wrong predictions and low accuracy results (such as 13% for *Lawsonia*, 28% for *Ca*, among others), drastically decreasing the final overall score.

4.6. Performance study

In order to assess the usability of the application in real world scenarios it is important to understand how it performs, not only in terms of accuracy (discussed in previous experiments), but also in terms of execution times. For this purpose a subset of the MSH WSD Data Set was compiled, containing 3680 citations (about 15% of total data set) from 18 different terms. This subset contains over 590 thousand concepts, identified on more than 750 thousand words present in over 31 thousand phrases. Table 4.12 shows detailed information about this subset. The total execution time was approximately 17 hours and 47 minutes which means that an average of 17 seconds was needed for each citation to be disambiguated[16]. The most critical phase of the disambiguation process is the graph construction phase, being the breadth first search algorithm the most time expensive part of the application. This can be considered a significant amount of time in real time interaction applications. On the other hand, in applications used to disambiguate large amounts of documents, where usually the user is not interacting in real time this time frame does not represent much.

Table 4.12: Performance study subset detailed information

	Citations	Sentences	Words	Concepts
Total	3 680	31 136	756 398	594 234
Average per Sentence		8,4609	24,2934	19,0851
Average per Citation			205,5429	161,4766

[16] Experiments were conducted on a server with 8 processing cores at 2.26GHz and 40GB of memory.

Chapter 5

Conclusions

In summary, there is still plenty of work to be done in word sense disambiguation field. It is an important problem of computer science that grows constantly as more and more information becomes available. The most popular approaches to solve this problem are based on supervised machine learning, where a classifier is trained on manually curated training instances to generate a model that can be used to classify future instances. However, such systems are limited to the amount of available training data.

In this thesis an approach for biomedical WSD problem, that does not rely on manually curated training data, is proposed. Such approaches have great importance because of the wider applicability, i.e., while supervised solutions are focused on small sub-domains of the biomedical field, knowledge-based approaches can easily address the entire domain. On the other hand, because the latter solutions are more general the results tend to be poorer if compared on the basis of a specific evaluation corpus.

Nevertheless, in a field such as biomedicine there are several advantages of using knowledge-based approaches. The fact that the domain is wide and the data is growing at a fast pace, allied to the fact that the various sub-domains of the field are often related to each other implies that general WSD solutions are helpful. It is also important to note that the word sense disambiguation problem in biomedical field is not an easy task, not even for curators sometimes.

Given the fact that the biomedical field is constantly evolving, also the biomedical tools need to constantly adapt and improve. For that reason, all work done in this thesis was thought to be adaptable, scalable and extensible. The most important phases of the presented tool are the knowledge-based graph creation and the scoring phase. Thanks to

a modular implementation, not only the knowledge source can easily be upgraded, but also this tool can easily use a different source of information to create graphs. Also, regarding the scoring phase, this tool can be equipped with more scoring algorithms that can be fine-tuned to improve the results of specific domains.

Overall, the proposed solution achieves satisfactory results even though more work can be done to improve its performance. To understand and identify cases where the lack of information impacts negatively the WSD task, and to be able to discover and provide complementary information for these cases is the key to solve the cases where the achieved accuracy was below average.

It is the aim of this thesis to provide a tool that will help in the task of biomedical research and aid the curators in their day-to-day tasks. Furthermore, the easy to integrate structure of the solution enables researchers to implement or upgrade their tools integrating a WSD component, which adds a lot of potential.

Chapter 6

Future Work

Some possible future research based on the work done for this thesis categorizes, mainly, into two groups: knowledge source research and scoring techniques research.

Based on state-of-the-art solutions it is noticeable that implementing and fine-tuning new scoring algorithms is of the utmost importance. The best results can be obtained when combining the output from multiple solutions. The proposed architecture in this thesis has the advantage to be scalable and extensible, facilitating the future expansion and improvement of the application. The use of graphs is also an important choice as there are several known algorithms that use graphs to perform WSD tasks.

Regarding the knowledge source, analyzing all performed experiments it is easy to understand that the more information the better the results will be. For that reason it is believed that discovering and integrating new sources of information can greatly improve the performance of the application.

In summary, the proposed application already proved to be successful in word sense disambiguation task and has great potential to be further developed taking as basis the work done so far.

Bibliography

[1] R. Blumberg and S. Atre, “The Problem with Unstructured Data.,” *DM Rev.*, vol. 13, no. 2, p. 42, 2003.

[2] U. M. Fayyad, G. Piatetsky-shapiro, and P. Smyth, “From data mining to knowledge discovery: an overview,” in *Advances in Knowledge Discovery and Data Mining*, vol. 17, no. 3, 1996, pp. 1–34.

[3] L. Hirschman, “The Evolution of evaluation: Lessons from the Message Understanding Conferences,” *Computer Speech & Language*, vol. 12, no. 4. pp. 281–305, 1998.

[4] A. J. Jimeno-Yepes and A. R. Aronson, “Knowledge-based biomedical word sense disambiguation: comparison of approaches.,” *BMC Bioinformatics*, vol. 11, no. 1, p. 569, Jan. 2010.

[5] M. Weeber, J. G. Mork, and a R. Aronson, “Developing a test collection for biomedical word sense disambiguation.,” *Proc. AMIA Symp.*, pp. 746–750, 2001.

[6] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” *Proc. AMIA Symp.*, pp. 17–21, 2001.

[7] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances.,” *J. Am. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, 2010.

[8] R. Navigli, “Word sense disambiguation: A survey,” *ACM Comput. Surv.*, vol. 41, no. 2, p. 10, 2009.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database*, vol. 71. 1998.

[10] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminology.,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D267–D270, 2004.

[11] A. Jimeno-Yepes and A. R. Aronson, “Self-training and co-training in biomedical word sense disambiguation,” in *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011*, 2011, pp. 182–183.

[12] H. Liu, V. Teller, and C. Friedman, “A multi-aspect comparison study of supervised word sense disambiguation,” *J. Am. Med. Informatics Assoc.*, vol. 11, pp. 320–331, 2004.

[13] E. Agirre and A. Soroa, “Semeval-2007 Task 02 : Evaluating Word Sense Induction and Discrimination Systems,” *Comput. Linguist.*, pp. 7–12, 2007.

[14] S. Manandhar and I. P. Klapaftis, “SemEval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems,” in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 2009, pp. 117–122.

[15] T. Martín-Wanton, R. Berlanga-Llavori, and A. Jimeno-Yepes, “Preliminary results for biomedical word sense disambiguation based on semantic clustering,” *Proc. - Int. Work. Database Expert Syst. Appl. DEXA*, pp. 460–464, 2011.

[16] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch, “Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 1, pp. 96–113, 2006.

[17] B. McInnes, “An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline,” *Proc. 46th Annu. Meet. …*, pp. 49–54, 2008.

[18] G. a. Miller, “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez, “Disambiguation of biomedical text using diverse sources of information.,” *BMC Bioinformatics*, vol. 9 Suppl 11, p. S7, 2008.

[20] H. Liu, S. B. Johnson, and C. Friedman, “Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS,” *J. Am. Med. Informatics Assoc.*, vol. 9, pp. 621–636, 2002.

[21] E. Agirre and G. Rigau, “Word Sense Disambiguation using Conceptual Density,” *Proc. 16th Conf. Comput. Linguist.*, vol. 1, p. 8, 1996.

[22] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification,” in *WordNet: An electronic lexical database.*, 1998, pp. 265–283.

[23] D. Lin, “Automatic retrieval and clustering of similar words,” in *ACL ’98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1998, pp. 768–774.

[24] P. Resnik, “Using information content to evaluate seantic similarity in a taxonomy,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[25] R. Mihalcea, “Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling,” *Proc. Conf. Hum. Lang. Technol. Empir. Methods Nat. Lang. Process.*, pp. 411–418, 2005.

[26] E. Agirre and A. Soroa, “Personalizing PageRank for Word Sense Disambiguation,” in *Proceedings of the 12th Conference of the European Chapter of the ACL*, 2009, pp. 33–41.

[27] B. T. McInnes, T. Pedersen, Y. Liu, G. B. Melton, and S. V Pakhomov, “Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity.,” *AMIA Annu. Symp. Proc.*, vol. 2011, pp. 895–904, 2011.

[28] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* -, 1994, pp. 133–138.

[29] H. A. Nguyen and H. Al-Mubaid, “New ontology-based semantic similarity measure for the biomedical domain,” *2006 IEEE Int. Conf. Granul. Comput.*, 2006.

[30] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” p. 15, Sep. 1997.

[31] D. Lin, “An Information-Theoretic Definition of Similarity,” in *Proceedings of ICML*, 1998, pp. 296–304.

[32] R. Navigli and M. Lapata, “An experimental study of graph connectivity for unsupervised word sense disambiguation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 678–692, 2010.

[33] T. Nunes, D. Campos, S. Matos, and J. L. Oliveira, “BeCAS: Biomedical concept recognition services and visualization,” *Bioinformatics*, vol. 29, no. 15, pp. 1915–1916, 2013.

[34] N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen, “Comparison of concept recognizers for building the Open Biomedical Annotator.,” *BMC Bioinformatics*, vol. 10 Suppl 9, p. S14, 2009.

[35] R. Kaushik, P. Bohannon, J. F. Naughton, and H. F. Korth, “Covering indexes for branching path queries,” *Proc. 2002 ACM SIGMOD Int. Conf. Manag. data - SIGMOD ’02*, p. 133, 2002.

[36] A. J. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, “Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation.,” *BMC Bioinformatics*, vol. 12, no. 1, p. 223, Jan. 2011.

[37] D. Campos, S. Matos, and J. L. Oliveira, “A modular framework for biomedical concept recognition.,” *BMC Bioinformatics*, vol. 14, no. 1, p. 281, Jan. 2013.

Appendix A

NLM-WSD Dataset

This appendix contains the sense distribution of the target words in the MSH WSD dataset along with their corresponding CUIs in the 2009AB UMLS.

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset

Word	M1	M2	M3	M4	M5	Total
AA	99	99	0	0	0	198
ADA	99	99	0	0	0	198
ADH	99	99	0	0	0	198
ADP	99	50	0	0	0	149
ALS	99	99	0	0	0	198
ANA	99	99	0	0	0	198
Adrenal	99	99	0	0	0	198
Ala	99	99	99	0	0	297
Arteriovenous Anastomoses	30	99	0	0	0	129
Astragalus	99	99	0	0	0	198
B-Cell Leukemia	92	66	0	0	0	158
BAT	99	99	0	0	0	198
BLM	99	99	0	0	0	198
BPD	99	99	0	0	0	198
BR	71	99	0	0	0	170
BSA	99	99	0	0	0	198
BSE	99	99	0	0	0	198
Borrelia	99	99	0	0	0	198
Brucella abortus	81	99	0	0	0	180
CAD	99	99	0	0	0	198
CAM	99	99	0	0	0	198
CCD	42	99	0	0	0	141

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5	Total
CCl4	99	99	0	0	0	198
CDA	99	99	0	0	0	198
CDR	48	99	0	0	0	147
CH	91	57	0	0	0	148
CIS	99	54	0	0	0	153
CI	84	99	0	0	0	183
CLS	17	17	0	0	0	34
CNS	99	99	0	0	0	198
CPDD	13	22	0	0	0	35
CP	99	99	99	0	0	297
CRF	99	99	0	0	0	198
CTX	84	99	0	0	0	183
Ca	99	99	99	99	0	396
Callus	99	51	0	0	0	150
Cardiac pacemaker	99	99	0	0	0	198
Cell	99	99	0	0	0	198
Cement	99	86	0	0	0	185
Cholera	99	99	0	0	0	198
Cilia	99	57	0	0	0	156
Coffee	99	99	0	0	0	198
Cold	99	62	99	0	0	260
Compliance	99	99	0	0	0	198
Cortex	99	99	0	0	0	198
Cortical	99	99	99	0	0	297
Crack	99	64	0	0	0	163
Crown	99	99	0	0	0	198
DAT	99	99	0	0	0	198
DBA	84	99	0	0	0	183
DDD	99	99	0	0	0	198
DDS	99	22	99	0	0	220
DE	27	99	0	0	0	126
DI	99	99	0	0	0	198
DON	27	99	0	0	0	126
Digestive	99	99	0	0	0	198
EGG	99	99	0	0	0	198
EMS	99	99	0	0	0	198
EM	99	30	0	0	0	129
ERP	99	99	0	0	0	198

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5	Total
ERUPTION	99	99	0	0	0	198
Eels	31	99	0	0	0	130
Epi	99	99	0	0	0	198
Erythrocytes	99	99	0	0	0	198
Exercises	99	99	0	0	0	198
FAS	99	99	0	0	0	198
FA	99	99	0	0	0	198
FTC	99	99	0	0	0	198
Familial Adenomatous Poly-posis	99	99	0	0	0	198
Fe	99	99	0	0	0	198
Fish	99	99	0	0	0	198
Follicle	99	99	0	0	0	198
Follicles	99	99	0	0	0	198
GAG	99	99	0	0	0	198
Gamma-Interferon	99	99	0	0	0	198
Ganglion	99	99	0	0	0	198
Gas	99	99	0	0	0	198
Glycoside	99	99	0	0	0	198
HCl	99	99	0	0	0	198
HGF	93	99	0	0	0	192
HHV 8	99	77	0	0	0	176
HIV	99	99	0	0	0	198
HPS	79	99	0	0	0	178
HR	10	99	0	0	0	109
Haemophilus ducreyi	54	99	0	0	0	153
Hemlock	57	20	0	0	0	77
Heregulin	74	99	0	0	0	173
Hip	99	66	0	0	0	165
Hybridization	99	99	0	0	0	198
IA	35	99	0	0	0	134
INDO	23	99	0	0	0	122
IP	97	99	0	0	0	196
ITP	99	99	0	0	0	198
Ice	37	99	99	0	0	235
Ion	99	99	0	0	0	198
Iris	99	62	0	0	0	161
JP	93	99	0	0	0	192

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5	Total
LABOR	99	99	0	0	0	198
Lactation	99	99	0	0	0	198
Language	99	99	0	0	0	198
Laryngeal	99	99	0	0	0	198
Lawsonia	16	99	0	0	0	115
Leishmaniasis	62	99	0	0	0	161
Lupus	99	99	99	0	0	297
MAF	99	21	0	0	0	120
MBP	99	44	0	0	0	143
MCC	32	99	0	0	0	131
MHC	99	99	0	0	0	198
MRS	67	99	0	0	0	166
Malaria	99	99	0	0	0	198
Medullary	99	99	0	0	0	198
Milk	99	99	0	0	0	198
Moles	99	75	0	0	0	174
Murine sarcoma virus	99	81	0	0	0	180
NBS	99	47	0	0	0	146
NEUROFIBROMATOSIS	99	99	0	0	0	198
NM	38	84	0	0	0	122
NPC	64	99	0	0	0	163
Nurse	99	99	0	0	0	198
Nursing	99	99	0	0	0	198
OCD	99	99	0	0	0	198
OH	99	99	0	0	0	198
ORI	99	24	0	0	0	123
Orf	99	99	0	0	0	198
PAC	16	46	0	0	0	62
PAF	16	99	0	0	0	115
PCA	99	95	99	99	99	491
PCB	99	28	0	0	0	127
PCD	99	99	0	0	0	198
PCP	99	99	99	0	0	297
PEP	99	99	0	0	0	198
PHA	99	11	0	0	0	110
POL	99	63	0	0	0	162
PR	66	99	0	0	0	165
PVC	99	99	0	0	0	198

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5	Total
Parotitis	99	99	0	0	0	198
Pharmaceutical	99	99	0	0	0	198
Phosphorus	99	99	0	0	0	198
Phosphorylase	99	67	0	0	0	166
Plague	99	69	0	0	0	168
Plaque	99	99	0	0	0	198
Platelet	99	99	0	0	0	198
Pleuropneumonia	99	99	0	0	0	198
Pneumocystis	99	99	0	0	0	198
Polymyalgia Rheumatica	99	99	0	0	0	198
Potassium	99	99	0	0	0	198
Projection	99	99	0	0	0	198
RA	99	99	99	0	0	297
RBC	99	99	0	0	0	198
RB	99	99	0	0	0	198
RSV	35	99	0	0	0	134
Radiation	99	99	0	0	0	198
Respiration	99	99	0	0	0	198
Retinal	99	99	0	0	0	198
Root	99	99	0	0	0	198
SARS-associated coronavirus	47	71	0	0	0	118
SARS	99	99	0	0	0	198
SCD	99	99	0	0	0	198
SLS	65	99	0	0	0	164
SPR	99	99	0	0	0	198
SS	98	46	0	0	0	144
STEM	99	99	0	0	0	198
Schistosoma mansoni	99	99	0	0	0	198
Semen	87	99	0	0	0	186
Sodium	98	99	0	0	0	197
Staph	99	99	0	0	0	198
Sterilization	99	99	0	0	0	198
Strep	99	98	0	0	0	197
Synapsis	35	99	0	0	0	134
TAT	99	99	99	0	0	297
TEM	99	99	0	0	0	198
THYMUS	99	99	99	0	0	297
TLC	99	99	0	0	0	198

Table A.1: Sense distribution for ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5	Total
TMJ	99	99	0	0	0	198
TMP	99	51	0	0	0	150
TNC	68	99	0	0	0	167
TNT	99	99	0	0	0	198
TPA	99	99	0	0	0	198
TPO	99	99	0	0	0	198
TRF	99	80	0	0	0	179
TSF	35	18	0	0	0	53
TYR	99	99	0	0	0	198
Tax	81	99	0	0	0	180
Tolerance	99	99	0	0	0	198
Torula	34	88	0	0	0	122
US	99	99	0	0	0	198
Ventricles	99	99	0	0	0	198
WBS	93	35	0	0	0	128
WT1	99	99	0	0	0	198
Wasp	99	99	0	0	0	198
Yellow Fever	99	84	0	0	0	183
cRNA	99	99	0	0	0	198
dC	99	99	0	0	0	198
drinking	99	99	0	0	0	198
eCG	99	99	0	0	0	198
lens	99	99	99	0	0	297
lymphogranulomatosis	20	99	0	0	0	119
pI	99	57	0	0	0	156
posterior pituitary	95	99	0	0	0	194
rDNA	99	99	0	0	0	198
sex factor	35	96	0	0	0	131
tomography	99	99	0	0	0	198
veterinary	99	99	0	0	0	198

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset

Word	M1	M2	M3	M4	M5
AA	C0002520	C0001972			
ADA	C0002456	C0001457			
ADH	C0001942	C0003779			
ADP	C0001459	C0004374			
Adrenal	C0014563	C0001625			
Ala	C0001898	C0002563	C0051405		
ALS	C0003372	C0002736			
ANA	C0002463	C0003243			
Arteriovenous Anastomoses	C0684204	C0225984			
Astragalus	C0039277	C0330845			
B-Cell Leukemia	C2004493	C0023434			
BAT	C0006298	C0008139			
BLM	C0005859	C0005740			
Borrelia	C0024198	C0006033			
BPD	C0006287	C0006012			
BR	C0006137	C0006222			
Brucella abortus	C0302363	C0006304			
BSA	C0005902	C0036774			
BSE	C0085209	C0085105			
Ca	C0006675	C0006754	C0019564	C0006823	
CAD	C0011905	C1956346			
Callus	C0006767	C0376154			
CAM	C0007578	C0178551			
Cardiac pacemaker	C0037189	C0030163			
CCD	C0751951	C0008928			
CCl4	C0209338	C0007022			
CDA	C0092801	C0002876			
CDR	C0011485	C0021024			
Cell	C0007634	C1136359			
Cement	C1706094	C0011343			
CH	C0008115	C0039021			
Cholera	C0008354	C0008359			
CI	C0022326	C0008107			
Cilia	C0008778	C0015422			
CIS	C0007099	C0162854			
CLS	C0265252	C0343084			
CNS	C0028654	C0927232			

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset

Word	M1	M2	M3	M4	M5
Coffee	C0085952	C0009237			
Cold	C0009264	C0024117	C0009443		
Compliance	C0009563	C1321605			
Cortex	C0007776	C0001614			
Cortical	C0022655	C0007776	C0001613		
CP	C0007789	C0033477	C0008925		
CPDD	C0553730	C0008838			
Crack	C0085163	C0040441			
CRF	C0022661	C0010132			
cRNA	C1321571	C0056208			
Crown	C0226993	C0010384			
CTX	C0238052	C0010583			
DAT	C0002395	C0114838			
DBA	C1260899	C0025923			
dC	C0012764	C0011485			
DDD	C0026256	C0011037			
DDS	C0010980	C0950121	C0085104		
DE	C0011198	C0017480			
DI	C0011848	C0032246			
Digestive	C0012240	C0012238			
DON	C0028652	C0012020			
drinking	C0684271	C0001948			
eCG	C0018064	C1623258			
Eels	C0677644	C0013671			
EGG	C0029974	C0013710			
EM	C0026019	C0014921			
EMS	C0013961	C0015063			
Epi	C0014563	C0014582			
ERP	C0015214	C0008310			
ERUPTION	C0015230	C1533692			
Erythrocytes	C0014792	C0014772			
Exercises	C0452240	C0015259			
FA	C0016410	C0015625			
Familial Adenomatous Polyposis	C0162832	C0032580			
FAS	C0015683	C0015923			
Fe	C0376520	C0302583			
Fish	C0016163	C0162789			

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5
Follicle	C0221971	C0018120			
Follicles	C0221971	C0018120			
FTC	C0041713	C0206682			
GAG	C0017346	C0017973			
Gamma-Interferon	C0021740	C0021745			
Ganglion	C0017067	C1258666			
Gas	C0017110	C0016204			
Glycoside	C0007158	C0017977			
Haemophilus ducreyi	C0007947	C0018481			
HCl	C0023443	C0020259			
Hemlock	C0949851	C0242872			
Heregulin	C0752253	C0626201			
HGF	C0021760	C0062534			
HHV 8	C0376526	C0036220			
Hip	C0019552	C0022122			
HIV	C0019693	C0019682			
HPS	C0242994	C0079504			
HR	C0010343	C0018810			
Hybridization	C0020202	C0028602			
IA	C0022037	C0021487			
Ice	C0025611	C0020746	C0534519		
INDO	C0021247	C0021246			
Ion	C0022024	C0022023			
IP	C0021069	C0021171			
Iris	C0022077	C1001362			
ITP	C0021540	C0043117			
JP	C0031106	C0022341			
LABOR	C0022864	C0043227			
Lactation	C0022925	C0006147			
Language	C0033348	C0023008			
Laryngeal	C0023078	C0023081			
Lawsonia	C1068388	C0752045			
Leishmaniasis	C1548483	C0023281			
lens	C0023318	C0023308	C0023317		
Lupus	C0024131	C0024141	C0024138		
lymphogranulomatosis	C0036202	C0019829			
MAF	C0079786	C0919482			
Malaria	C0206255	C0024530			

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5
MBP	C0014063	C0065661			
MCC	C0162804	C0007129			
Medullary	C0001629	C0025148			
MHC	C0027100	C0024518			
Milk	C0026131	C0026140			
Moles	C0324740	C0027960			
MRS	C0025235	C0024487			
Murine sarcoma virus	C0026630	C0026399			
NBS	C0027819	C0398791			
NEUROFIBROMATOSIS	C0162678	C0085113			
NM	C0025033	C0027972			
NPC	C0220756	C0028587			
Nurse	C0028661	C0006147			
Nursing	C0028677	C0006147			
OCD	C0028768	C0029421			
OH	C0063146	C0028905			
Orf	C0079941	C0013570			
ORI	C0242961	C0206601			
PAC	C0033036	C0949780			
PAF	C0037019	C0032172			
Parotitis	C0026780	C0030583			
PCA	C0429865	C0149576	C0078944	C0030625	C0030131
PCB	C0032447	C0033223			
PCD	C0162638	C0022521			
PCP	C0032305	C0030855	C0031381		
PEP	C0031642	C0135981			
PHA	C0031858	C0030779			
Pharmaceutical	C0013058	C0031336			
Phosphorus	C0080014	C0031705			
Phosphorylase	C0017916	C0917783			
pI	C0022171	C0812425			
Plague	C0032064	C0032066			
Plaque	C0333463	C0011389			
Platelet	C0032181	C0005821			
Pleuropneumonia	C0032241	C0026934			
Pneumocystis	C0032305	C0597258			
POL	C0017360	C0032356			
Polymyalgia Rheumatica	C0032533	C0039483			

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5
posterior pituitary	C0032017	C0032009			
Potassium	C0032821	C0162800			
PR	C0034044	C0034833			
Projection	C0016538	C0033363			
PVC	C0151636	C0032624			
RA	C0002893	C0034625	C0003873		
Radiation	C0851346	C1522449			
RB	C0035335	C0035930			
RBC	C0014792	C0014772			
rDNA	C0012931	C0012933			
Respiration	C0035203	C0282636			
Retinal	C0035331	C0035298			
Root	C0242726	C0040452			
RSV	C0086943	C0035236			
SARS-associated coronavirus	C1175175	C1175743			
SARS	C1175175	C1175743			
SCD	C0085298	C0002895			
Schistosoma mansoni	C0036319	C0036330			
Semen	C0036563	C0036614			
sex factor	C0036881	C0015435			
SLS	C0037231	C0037506			
Sodium	C0037473	C0037570			
SPR	C0164209	C0597731			
SS	C0039101	C0085077			
Staph	C0038160	C0038170			
STEM	C0242767	C0162731			
Sterilization	C0038288	C0038280			
Strep	C0038402	C0038395			
Synapsis	C0039062	C0598501			
TAT	C0039756	C0017375	C0039341		
Tax	C0144576	C0039371			
TEM	C0678118	C0040975			
THYMUS	C0040112	C0040113	C1015036		
TLC	C0040509	C0008569			
TMJ	C0039496	C0039493			
TMP	C0041041	C0040079			
TNC	C0076088	C0077400			

Table A.2: Possible CUIs for the ambiguous terms in MSH WSD dataset (Cont.)

Word	M1	M2	M3	M4	M5
TNT	C0077404	C0041070			
Tolerance	C0013220	C0020963			
tomography	C0040395	C0040405			
Torula	C0010414	C0010415			
TPA	C0039654	C0032143			
TPO	C0040052	C0021965			
TRF	C0040162	C0021759			
TSF	C0040052	C0021756			
TYR	C0041485	C0041484			
US	C0041703	C0041618			
Ventricles	C0018827	C0007799			
veterinary	C0206212	C0042615			
Wasp	C0258432	C0043041			
WBS	C0175702	C0004903			
WT1	C0148873	C0027708			
Yellow Fever	C0043395	C0301508			

Table A.3: Semantic types frequency count in MSH WSD dataset

TUI	Full Semantic Type Name	# Occurences
T047	Disease or Syndrome	73
T116	Amino Acid, Peptide, or Protein	48
T121	Pharmacologic Substance	44
T123	Biologically Active Substance	30
T023	Body Part, Organ, or Organ Component	26
T109	Organic Chemical	25
T083	Geographic Area	20
T129	Immunologic Factor	17
T191	Neoplastic Process	15
T019	Congenital Abnormality	10
T126	Enzyme	10
T131	Hazardous or Poisonous Substance	10
T114	Nucleic Acid, Nucleoside, or Nucleotide	10
T002	Plant	10
T060	Diagnostic Procedure	9
T196	Element, Ion, or Isotope	9
T059	Laboratory Procedure	9
T007	Bacterium	8
T028	Gene or Genome	8
T197	Inorganic Chemical	8
T125	Hormone	7
T061	Therapeutic or Preventive Procedure	7
T005	Virus	7
T074	Medical Device	5
T097	Professional or Occupational Group	5
T081	Quantitative Concept	5
T122	Biomedical or Dental Material	4
T030	Body Space or Junction	4
T025	Cell	4
T168	Food	4
T040	Organism Function	4
T020	Acquired Abnormality	3
T195	Antibiotic	3
T091	Biomedical Occupation or Discipline	3
T031	Body Substance	3
T118	Carbohydrate	3
T043	Cell Function	3
T119	Lipid	3

Table A.3: Semantic types frequency count in MSH WSD dataset (Cont.)

TUI	Full Semantic Type Name	# Occurences
T015	Mammal	3
T070	Natural Phenomenon or Process	3
T124	Neuroreactive Substance or Biogenic Amine	3
T042	Organ or Tissue Function	3
T046	Pathologic Function	3
T022	Body System	2
T026	Cell Component	2
T204	Eukaryote	2
T033	Finding	2
T013	Fish	2
T004	Fungus	2
T130	Indicator, Reagent, or Diagnostic Aid	2
T055	Individual Behavior	2
T170	Intellectual Product	2
T048	Mental or Behavioral Dysfunction	2
T063	Molecular Biology Research Technique	2
T115	Organophosphorus Compound	2
T094	Professional Society	2
T192	Receptor	2
T024	Tissue	2
T127	Vitamin	2
T104	Chemical Viewed Structurally	1
T201	Clinical Attribute	1
T056	Daily or Recreational Activity	1
T018	Embryonic Structure	1
T169	Functional Concept	1
T045	Genetic Function	1
T102	Group Attribute	1
T058	Health Care Activity	1
T093	Health Care Related Organization	1
T037	Injury or Poisoning	1
T034	Laboratory or Test Result	1
T171	Language	1
T066	Machine Activity	1
T073	Manufactured Object	1
T041	Mental Process	1
T057	Occupational Activity	1
T032	Organism Attribute	1

Table A.3: Semantic types frequency count in MSH WSD dataset (Cont.)

TUI	Full Semantic Type Name	# Occurences
T092	Organization	1
T039	Physiologic Function	1
T075	Research Device	1
T095	Self-help or Relief Organization	1
T184	Sign or Symptom	1
T110	Steroid	1

Table A.4: Uncovered semantic types by Neji dictionaries

TUI	Full Semantic Type Name	# Occurrences	# Covered
T083	Geographic Area	20	0
T060	Diagnostic Procedure	9	0
T059	Laboratory Procedure	9	0
T061	Therapeutic or Preventive Procedure	7	0
T074	Medical Device	5	0
T097	Professional or Occupational Group	5	0
T081	Quantitative Concept	5	0
T122	Biomedical or Dental Material	4	0
T168	Food	4	0
T091	Biomedical Occupation or Discipline	3	0
T070	Natural Phenomenon or Process	3	0
T055	Individual Behavior	2	0
T170	Intellectual Product	2	0
T063	Molecular Biology Research Technique	2	0
T094	Professional Society	2	0
T201	Clinical Attribute	1	0
T056	Daily or Recreational Activity	1	0
T169	Functional Concept	1	0
T102	Group Attribute	1	0
T058	Health Care Activity	1	0
T093	Health Care Related Organization	1	0
T034	Laboratory or Test Result	1	0
T171	Language	1	0
T066	Machine Activity	1	0
T073	Manufactured Object	1	0
T041	Mental Process	1	0
T057	Occupational Activity	1	0
T032	Organism Attribute	1	0
T092	Organization	1	0
T075	Research Device	1	0
T095	Self-help or Relief Organization	1	0

Appendix B

Experiments detailed results

Table B.1: Detailed results of UMLS experiment

Term	Total Instances	Correct Predictions	Score
ADH	198	120	0,6061
Adrenal	198	120	0,6061
Ala	297	116	0,3906
ALS	196	109	0,5561
Arteriovenous Anastomoses	129	86	0,6667
Astragalus	195	102	0,5231
BAT	198	102	0,5152
B-Cell Leukemia	158	66	0,4177
BLM	198	94	0,4747
Borrelia	198	97	0,4899
BPD	198	126	0,6364
Ca	396	118	0,2980
Callus	150	62	0,4133
CAM	198	142	0,7172
CCD	141	124	0,8794
CCl4	198	100	0,5051
CDA	198	142	0,7172
CDR	147	114	0,7755
CLS	34	34	1,0000
Cold	259	92	0,3552
Cortex	198	163	0,8232
Cortical	297	235	0,7912
CP	297	188	0,6330
CPDD	35	24	0,6857
Crack	163	116	0,7117

Table B.1: Detailed results of UMLS experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
CRF	198	176	0,8889
CTX	183	125	0,6831
DAT	198	136	0,6869
DBA	183	84	0,4590
DDD	198	103	0,5202
DDS	219	74	0,3379
Epi	198	162	0,8182
ERUPTION	198	99	0,5000
FA	198	149	0,7525
Familial Adenomatous Polyposis	198	101	0,5101
FAS	198	121	0,6111
Fe	198	90	0,4545
Follicle	198	190	0,9596
Follicles	198	177	0,8939
GAG	198	101	0,5101
Gamma-Interferon	198	99	0,5000
Ganglion	198	99	0,5000
Gas	197	95	0,4822
Glycoside	198	180	0,9091
Haemophilus ducreyi	153	62	0,4052
HCl	198	145	0,7323
Hemlock	77	20	0,2597
Heregulin	173	90	0,5202
HGF	192	100	0,5208
HHV 8	172	73	0,4244
Hip	165	77	0,4667
HIV	198	107	0,5404
HPS	178	163	0,9157
Ice	235	118	0,5021
Ion	198	98	0,4949
Iris	161	99	0,6149
ITP	198	99	0,5000
Lactation	198	118	0,5960
Lawsonia	115	15	0,1304
Leishmaniasis	161	83	0,5155
lens	295	99	0,3356
Lupus	297	132	0,4444

Table B.1: Detailed results of UMLS experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
lymphogranulomatosis	119	65	0,5462
MAF	119	45	0,3782
Malaria	198	99	0,5000
MBP	143	94	0,6573
MCC	131	100	0,7634
Medullary	198	158	0,7980
MHC	198	126	0,6364
Milk	197	115	0,5838
Moles	171	72	0,4211
Murine sarcoma virus	180	99	0,5500
NBS	146	104	0,7123
NEUROFIBROMATOSIS	197	98	0,4975
NPC	163	133	0,8160
OCD	198	176	0,8889
Orf	198	110	0,5556
PAF	115	71	0,6174
Parotitis	193	96	0,4974
PCA	491	185	0,3768
PCB	127	95	0,7480
PCD	198	188	0,9495
PCP	297	109	0,3670
PEP	198	153	0,7727
PHA	110	63	0,5727
Phosphorus	198	91	0,4596
Phosphorylase	166	93	0,5602
Plague	168	122	0,7262
Plaque	198	101	0,5101
Pleuropneumonia	198	106	0,5354
Pneumocystis	198	82	0,4141
Polymyalgia Rheumatica	198	119	0,6010
posterior pituitary	194	126	0,6495
RA	297	94	0,3165
RB	198	106	0,5354
rDNA	198	92	0,4646
Respiration	198	153	0,7727
Retinal	198	127	0,6414
Root	198	124	0,6263

Table B.1: Detailed results of UMLS experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
RSV	134	99	0,7388
SARS	172	110	0,6395
SARS-associated coronavirus	118	56	0,4746
SCD	198	128	0,6465
Schistosoma mansoni	196	114	0,5816
Semen	186	98	0,5269
sex factor	131	90	0,6870
SLS	164	110	0,6707
SS	144	120	0,8333
Staph	198	117	0,5909
Sterilization	197	98	0,4975
Strep	197	96	0,4873
Synapsis	134	38	0,2836
THYMUS	297	134	0,4512
TMJ	198	120	0,6061
TMP	150	117	0,7800
TNC	167	90	0,5389
TNT	198	100	0,5051
Tolerance	198	127	0,6414
Torula	122	56	0,4590
TPA	198	99	0,5000
TPO	198	132	0,6667
TRF	179	162	0,9050
TSF	53	45	0,8491
TYR	191	126	0,6597
Ventricles	198	175	0,8838
Wasp	198	101	0,5101
WBS	128	113	0,8828
WT1	198	101	0,5101
Yellow Fever	181	124	0,6851

Table B.2: Detailed results of Most Frequent Sense experiment

Term	Total Instances	Correct Predictions	Score
ADH	198	114	0,5758
Adrenal	198	120	0,6061
Ala	297	109	0,3670
ALS	196	109	0,5561
Arteriovenous Anastomoses	129	84	0,6512
Astragalus	195	97	0,4974
BAT	198	130	0,6566
B-Cell Leukemia	158	66	0,4177
BLM	198	88	0,4444
Borrelia	198	99	0,5000
BPD	198	169	0,8535
Ca	396	118	0,2980
Callus	150	72	0,4800
CAM	198	140	0,7071
CCD	141	121	0,8582
CCl4	198	100	0,5051
CDA	198	145	0,7323
CDR	147	105	0,7143
CLS	34	33	0,9706
Cold	259	90	0,3475
Cortex	198	158	0,7980
Cortical	297	250	0,8418
CP	297	188	0,6330
CPDD	35	28	0,8000
Crack	163	115	0,7055
CRF	198	175	0,8838
CTX	183	119	0,6503
DAT	198	134	0,6768
DBA	183	85	0,4645
DDD	198	141	0,7121
DDS	219	79	0,3607
Epi	198	159	0,8030
ERUPTION	198	99	0,5000
FA	198	150	0,7576
Familial Adenomatous Polyposis	198	107	0,5404
FAS	198	120	0,6061
Fe	198	99	0,5000

Table B.2: Detailed results of Most Frequent Sense experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
Follicle	198	191	0,9646
Follicles	198	190	0,9596
GAG	198	100	0,5051
Gamma-Interferon	198	99	0,5000
Ganglion	198	99	0,5000
Gas	197	95	0,4822
Glycoside	198	157	0,7929
Haemophilus ducreyi	153	62	0,4052
HCl	198	151	0,7626
Hemlock	77	40	0,5195
Heregulin	173	92	0,5318
HGF	192	98	0,5104
HHV 8	172	74	0,4302
Hip	165	87	0,5273
HIV	198	107	0,5404
HPS	178	157	0,8820
Ice	235	123	0,5234
Ion	198	99	0,5000
Iris	161	99	0,6149
ITP	198	101	0,5101
Lactation	198	117	0,5909
Lawsonia	115	33	0,2870
Leishmaniasis	161	102	0,6335
lens	295	98	0,3322
Lupus	297	127	0,4276
lymphogranulomatosis	119	84	0,7059
MAF	119	50	0,4202
Malaria	198	125	0,6313
MBP	143	91	0,6364
MCC	131	100	0,7634
Medullary	198	169	0,8535
MHC	198	127	0,6414
Milk	197	114	0,5787
Moles	171	72	0,4211
Murine sarcoma virus	180	99	0,5500
NBS	146	102	0,6986
NEUROFIBROMATOSIS	197	98	0,4975

Table B.2: Detailed results of Most Frequent Sense experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
NPC	163	133	0,8160
OCD	198	165	0,8333
Orf	198	114	0,5758
PAF	115	82	0,7130
Parotitis	193	96	0,4974
PCA	491	185	0,3768
PCB	127	114	0,8976
PCD	198	188	0,9495
PCP	297	109	0,3670
PEP	198	156	0,7879
PHA	110	68	0,6182
Phosphorus	198	99	0,5000
Phosphorylase	166	101	0,6084
Plague	168	125	0,7440
Plaque	198	100	0,5051
Pleuropneumonia	198	100	0,5051
Pneumocystis	198	82	0,4141
Polymyalgia Rheumatica	198	100	0,5051
posterior pituitary	194	124	0,6392
RA	297	101	0,3401
RB	198	106	0,5354
rDNA	198	110	0,5556
Respiration	198	158	0,7980
Retinal	198	125	0,6313
Root	198	142	0,7172
RSV	134	99	0,7388
SARS	172	107	0,6221
SARS-associated coronavirus	118	55	0,4661
SCD	198	128	0,6465
Schistosoma mansoni	196	116	0,5918
Semen	186	98	0,5269
sex factor	131	92	0,7023
SLS	164	114	0,6951
SS	144	117	0,8125
Staph	198	108	0,5455
Sterilization	197	98	0,4975
Strep	197	92	0,4670

Table B.2: Detailed results of Most Frequent Sense experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
Synapsis	134	36	0,2687
THYMUS	297	129	0,4343
TMJ	198	119	0,6010
TMP	150	110	0,7333
TNC	167	82	0,4910
TNT	198	100	0,5051
Tolerance	198	132	0,6667
Torula	122	44	0,3607
TPA	198	104	0,5253
TPO	198	135	0,6818
TRF	179	163	0,9106
TSF	53	45	0,8491
TYR	191	126	0,6597
Ventricles	198	171	0,8636
Wasp	198	107	0,5404
WBS	128	115	0,8984
WT1	198	104	0,5253
Yellow Fever	181	129	0,7127

Table B.3: Detailed results of Most Frequent Sense with acronym disambiguation

Term	Total Instances	Correct Predictions	Score
ADH	198	114	0,5758
Adrenal	198	120	0,6061
Ala	297	109	0,3670
ALS	196	113	0,5765
Arteriovenous Anastomoses	129	84	0,6512
Astragalus	195	97	0,4974
BAT	198	130	0,6566
B-Cell Leukemia	158	66	0,4177
BLM	198	88	0,4444
Borrelia	198	99	0,5000
BPD	198	196	0,9899

Table B.3: Detailed results of Most Frequent Sense with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
Ca	396	118	0,2980
Callus	150	72	0,4800
CAM	198	140	0,7071
CCD	141	140	0,9929
CCl4	198	100	0,5051
CDA	198	180	0,9091
CDR	147	105	0,7143
CLS	34	33	0,9706
Cold	259	90	0,3475
Cortex	198	158	0,7980
Cortical	297	250	0,8418
CP	297	188	0,6330
CPDD	35	26	0,7429
Crack	163	115	0,7055
CRF	198	176	0,8889
CTX	183	119	0,6503
DAT	198	139	0,7020
DBA	183	85	0,4645
DDD	198	141	0,7121
DDS	219	80	0,3653
Epi	198	159	0,8030
ERUPTION	198	99	0,5000
FA	198	150	0,7576
Familial Adenomatous Polyposis	198	107	0,5404
FAS	198	123	0,6212
Fe	198	99	0,5000
Follicle	198	191	0,9646
Follicles	198	190	0,9596
GAG	198	100	0,5051
Gamma-Interferon	198	99	0,5000
Ganglion	198	99	0,5000
Gas	197	95	0,4822
Glycoside	198	157	0,7929
Haemophilus ducreyi	153	62	0,4052
HCl	198	161	0,8131
Hemlock	77	40	0,5195
Heregulin	173	92	0,5318

Table B.3: Detailed results of Most Frequent Sense with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
HGF	192	98	0,5104
HHV 8	172	74	0,4302
Hip	165	87	0,5273
HIV	198	107	0,5404
HPS	178	157	0,8820
Ice	235	123	0,5234
Ion	198	99	0,5000
Iris	161	99	0,6149
ITP	198	162	0,8182
Lactation	198	117	0,5909
Lawsonia	115	33	0,2870
Leishmaniasis	161	102	0,6335
lens	295	98	0,3322
Lupus	297	127	0,4276
lymphogranulomatosis	119	84	0,7059
MAF	119	50	0,4202
Malaria	198	125	0,6313
MBP	143	91	0,6364
MCC	131	100	0,7634
Medullary	198	169	0,8535
MHC	198	127	0,6414
Milk	197	114	0,5787
Moles	171	72	0,4211
Murine sarcoma virus	180	99	0,5500
NBS	146	102	0,6986
NEUROFIBROMATOSIS	197	98	0,4975
NPC	163	141	0,8650
OCD	198	165	0,8333
Orf	198	114	0,5758
PAF	115	82	0,7130
Parotitis	193	96	0,4974
PCA	491	185	0,3768
PCB	127	114	0,8976
PCD	198	188	0,9495
PCP	297	139	0,4680
PEP	198	156	0,7879
PHA	110	68	0,6182

Table B.3: Detailed results of Most Frequent Sense with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
Phosphorus	198	99	0,5000
Phosphorylase	166	101	0,6084
Plague	168	125	0,7440
Plaque	198	100	0,5051
Pleuropneumonia	198	100	0,5051
Pneumocystis	198	82	0,4141
Polymyalgia Rheumatica	198	100	0,5051
posterior pituitary	194	124	0,6392
RA	297	101	0,3401
RB	198	114	0,5758
rDNA	198	110	0,5556
Respiration	198	158	0,7980
Retinal	198	125	0,6313
Root	198	142	0,7172
RSV	134	99	0,7388
SARS	172	107	0,6221
SARS-associated coronavirus	118	55	0,4661
SCD	198	128	0,6465
Schistosoma mansoni	196	116	0,5918
Semen	186	98	0,5269
sex factor	131	92	0,7023
SLS	164	114	0,6951
SS	144	126	0,8750
Staph	198	108	0,5455
Sterilization	197	98	0,4975
Strep	197	92	0,4670
Synapsis	134	36	0,2687
THYMUS	297	129	0,4343
TMJ	198	118	0,5960
TMP	150	111	0,7400
TNC	167	82	0,4910
TNT	198	100	0,5051
Tolerance	198	132	0,6667
Torula	122	44	0,3607
TPA	198	104	0,5253
TPO	198	136	0,6869
TRF	179	163	0,9106

Table B.3: Detailed results of Most Frequent Sense with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
TSF	53	50	0,9434
TYR	191	126	0,6597
Ventricles	198	171	0,8636
Wasp	198	107	0,5404
WBS	128	115	0,8984
WT1	198	104	0,5253
Yellow Fever	181	129	0,7127

Table B.4: Detailed results of MeSH Terms experiment

Term	Total Instances	Correct Predictions	Score
ADH	198	154	0,7778
Adrenal	198	121	0,6111
Ala	297	116	0,3906
ALS	196	117	0,5969
Arteriovenous Anastomoses	129	98	0,7597
Astragalus	195	103	0,5282
BAT	198	106	0,5354
B-Cell Leukemia	158	66	0,4177
BLM	198	102	0,5152
Borrelia	198	99	0,5000
BPD	198	193	0,9747
Ca	396	112	0,2828
Callus	150	91	0,6067
CAM	198	154	0,7778
CCD	141	130	0,9220
CCl4	198	100	0,5051
CDA	198	142	0,7172
CDR	147	117	0,7959
CLS	34	34	1,0000
Cold	259	92	0,3552
Cortex	198	170	0,8586
Cortical	297	243	0,8182

Table B.4: Detailed results of MeSH Terms experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
CP	297	196	0,6599
CPDD	35	29	0,8286
Crack	163	139	0,8528
CRF	198	184	0,9293
CTX	183	127	0,6940
DAT	198	142	0,7172
DBA	183	100	0,5464
DDD	198	130	0,6566
DDS	219	111	0,5068
Epi	198	162	0,8182
ERUPTION	198	99	0,5000
FA	198	152	0,7677
Familial Adenomatous Polyposis	198	104	0,5253
FAS	198	122	0,6162
Fe	198	99	0,5000
Follicle	198	189	0,9545
Follicles	198	177	0,8939
GAG	198	115	0,5808
Gamma-Interferon	198	99	0,5000
Ganglion	198	99	0,5000
Gas	197	97	0,4924
Glycoside	198	190	0,9596
Haemophilus ducreyi	153	64	0,4183
HCl	198	151	0,7626
Hemlock	77	32	0,4156
Heregulin	173	89	0,5145
HGF	192	110	0,5729
HHV 8	172	73	0,4244
Hip	165	101	0,6121
HIV	198	107	0,5404
HPS	178	171	0,9607
Ice	235	121	0,5149
Ion	198	104	0,5253
Iris	161	99	0,6149
ITP	198	99	0,5000
Lactation	198	118	0,5960
Lawsonia	115	15	0,1304

Table B.4: Detailed results of MeSH Terms experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
Leishmaniasis	161	100	0,6211
lens	295	100	0,3390
Lupus	297	143	0,4815
lymphogranulomatosis	119	101	0,8487
MAF	119	84	0,7059
Malaria	198	106	0,5354
MBP	143	108	0,7552
MCC	131	100	0,7634
Medullary	198	164	0,8283
MHC	198	147	0,7424
Milk	197	128	0,6497
Moles	171	72	0,4211
Murine sarcoma virus	180	109	0,6056
NBS	146	104	0,7123
NEUROFIBROMATOSIS	197	99	0,5025
NPC	163	133	0,8160
OCD	198	193	0,9747
Orf	198	134	0,6768
PAF	115	94	0,8174
Parotitis	193	112	0,5803
PCA	491	204	0,4155
PCB	127	122	0,9606
PCD	198	188	0,9495
PCP	297	108	0,3636
PEP	198	173	0,8737
PHA	110	73	0,6636
Phosphorus	198	116	0,5859
Phosphorylase	166	101	0,6084
Plague	168	112	0,6667
Plaque	198	101	0,5101
Pleuropneumonia	198	113	0,5707
Pneumocystis	198	83	0,4192
Polymyalgia Rheumatica	198	104	0,5253
posterior pituitary	194	131	0,6753
RA	297	109	0,3670
RB	198	106	0,5354
rDNA	198	126	0,6364

Table B.4: Detailed results of MeSH Terms experiment (Cont.)

Term	Total Instances	Correct Predictions	Score
Respiration	198	180	0,9091
Retinal	198	132	0,6667
Root	198	124	0,6263
RSV	134	99	0,7388
SARS	172	110	0,6395
SARS-associated coronavirus	118	57	0,4831
SCD	198	128	0,6465
Schistosoma mansoni	196	115	0,5867
Semen	186	98	0,5269
sex factor	131	90	0,6870
SLS	164	116	0,7073
SS	144	131	0,9097
Staph	198	121	0,6111
Sterilization	197	98	0,4975
Strep	197	102	0,5178
Synapsis	134	44	0,3284
THYMUS	297	147	0,4949
TMJ	198	121	0,6111
TMP	150	120	0,8000
TNC	167	148	0,8862
TNT	198	101	0,5101
Tolerance	198	126	0,6364
Torula	122	82	0,6721
TPA	198	99	0,5000
TPO	198	134	0,6768
TRF	179	163	0,9106
TSF	53	48	0,9057
TYR	191	153	0,8010
Ventricles	198	185	0,9343
Wasp	198	101	0,5101
WBS	128	121	0,9453
WT1	198	97	0,4899
Yellow Fever	181	130	0,7182

Table B.5: Detailed results of MeSH Terms with acronym disambiguation

Term	Total Instances	Correct Predictions	Score
ADH	198	154	0,7778
Adrenal	198	121	0,6111
Ala	297	116	0,3906
ALS	196	121	0,6173
Arteriovenous Anastomoses	129	98	0,7597
Astragalus	195	103	0,5282
BAT	198	104	0,5253
B-Cell Leukemia	158	66	0,4177
BLM	198	102	0,5152
Borrelia	198	99	0,5000
BPD	198	198	1,0000
Ca	396	112	0,2828
Callus	150	91	0,6067
CAM	198	154	0,7778
CCD	141	140	0,9929
CCl4	198	100	0,5051
CDA	198	180	0,9091
CDR	147	117	0,7959
CLS	34	34	1,0000
Cold	259	92	0,3552
Cortex	198	170	0,8586
Cortical	297	243	0,8182
CP	297	196	0,6599
CPDD	35	27	0,7714
Crack	163	139	0,8528
CRF	198	184	0,9293
CTX	183	127	0,6940
DAT	198	146	0,7374
DBA	183	100	0,5464
DDD	198	131	0,6616
DDS	219	111	0,5068
Epi	198	162	0,8182
ERUPTION	198	99	0,5000
FA	198	152	0,7677
Familial Adenomatous Polyposis	198	104	0,5253
FAS	198	123	0,6212
Fe	198	99	0,5000

Table B.5: Detailed results of MeSH Terms with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
Follicle	198	189	0,9545
Follicles	198	177	0,8939
GAG	198	115	0,5808
Gamma-Interferon	198	99	0,5000
Ganglion	198	99	0,5000
Gas	197	97	0,4924
Glycoside	198	190	0,9596
Haemophilus ducreyi	153	64	0,4183
HCl	198	154	0,7778
Hemlock	77	32	0,4156
Heregulin	173	89	0,5145
HGF	192	110	0,5729
HHV 8	172	76	0,4419
Hip	165	101	0,6121
HIV	198	107	0,5404
HPS	178	171	0,9607
Ice	235	121	0,5149
Ion	198	104	0,5253
Iris	161	99	0,6149
ITP	198	161	0,8131
Lactation	198	118	0,5960
Lawsonia	115	15	0,1304
Leishmaniasis	161	100	0,6211
lens	295	100	0,3390
Lupus	297	143	0,4815
lymphogranulomatosis	119	101	0,8487
MAF	119	84	0,7059
Malaria	198	106	0,5354
MBP	143	108	0,7552
MCC	131	100	0,7634
Medullary	198	164	0,8283
MHC	198	147	0,7424
Milk	197	128	0,6497
Moles	171	72	0,4211
Murine sarcoma virus	180	109	0,6056
NBS	146	104	0,7123
NEUROFIBROMATOSIS	197	99	0,5025

Table B.5: Detailed results of MeSH Terms with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
NPC	163	141	0,8650
OCD	198	193	0,9747
Orf	198	134	0,6768
PAF	115	94	0,8174
Parotitis	193	112	0,5803
PCA	491	204	0,4155
PCB	127	122	0,9606
PCD	198	188	0,9495
PCP	297	139	0,4680
PEP	198	173	0,8737
PHA	110	73	0,6636
Phosphorus	198	116	0,5859
Phosphorylase	166	101	0,6084
Plague	168	112	0,6667
Plaque	198	101	0,5101
Pleuropneumonia	198	113	0,5707
Pneumocystis	198	83	0,4192
Polymyalgia Rheumatica	198	104	0,5253
posterior pituitary	194	131	0,6753
RA	297	109	0,3670
RB	198	117	0,5909
rDNA	198	126	0,6364
Respiration	198	180	0,9091
Retinal	198	132	0,6667
Root	198	124	0,6263
RSV	134	99	0,7388
SARS	172	112	0,6512
SARS-associated coronavirus	118	57	0,4831
SCD	198	128	0,6465
Schistosoma mansoni	196	115	0,5867
Semen	186	98	0,5269
sex factor	131	90	0,6870
SLS	164	116	0,7073
SS	144	135	0,9375
Staph	198	121	0,6111
Sterilization	197	98	0,4975
Strep	197	102	0,5178

Table B.5: Detailed results of MeSH Terms with acronym disambiguation (Cont.)

Term	Total Instances	Correct Predictions	Score
Synapsis	134	44	0,3284
THYMUS	297	147	0,4949
TMJ	198	121	0,6111
TMP	150	120	0,8000
TNC	167	148	0,8862
TNT	198	101	0,5101
Tolerance	198	126	0,6364
Torula	122	82	0,6721
TPA	198	99	0,5000
TPO	198	135	0,6818
TRF	179	163	0,9106
TSF	53	50	0,9434
TYR	191	153	0,8010
Ventricles	198	185	0,9343
Wasp	198	101	0,5101
WBS	128	121	0,9453
WT1	198	97	0,4899
Yellow Fever	181	130	0,7182